PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G10L 5/02, 3/00, 3/02, 7/02

A1

(11) International Publication Number: **WO 98/01848**

(43) International Publication Date: 15 January 1998 (15.01.98)

(21) International Application Number: PCT/GB97/01831

(22) International Filing Date: 7 July 1997 (07.07.97)

(30) Priority Data:

9614209.6	5 July 1996 (05.07.96)	GB
021,815	16 July 1996 (16.07.96)	US

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

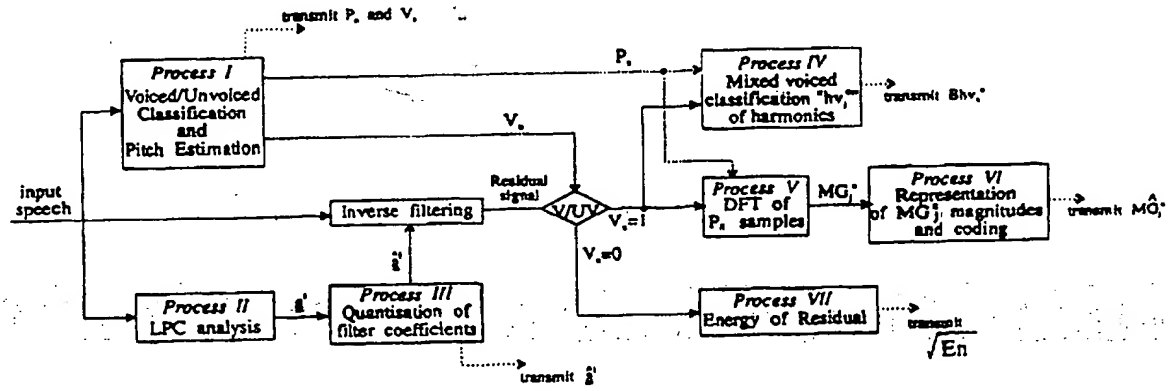
(71) Applicant *(for all designated States except US)*: THE VICTORIA UNIVERSITY OF MANCHESTER [GB/GB]; Oxford Road, Manchester M13 9PL (GB).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: XYDEAS, Costas [GB/GB]; 13 Thorngrove Hill, Wilmslow, Cheshire SK9 1DF (GB).

(74) Agent: ALLMAN, Peter, John; Marks & Clerk, Sussex House, 83-85 Mosley Street, Manchester M2 3LG (GB).

(54) Title: SPEECH SYNTHESIS SYSTEM



(57) Abstract

A speech synthesis system in which a speech signal is divided into a series of frames, and each frame is converted into a coded signal including a voiced/unvoiced classification and a pitch estimate, wherein a low pass filtered speech segment centred about a reference sample is defined in each frame, a correlation value is calculated for each of a series of candidate pitch estimates as the maximum of multiple crosscorrelation values obtained from variable length speech segments centred about the reference sample, the correlation values are used to form a correlation function defining peaks, and the locations of the peaks are determined and used to define a pitch estimate.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

SPEECH SYNTHESIS SYSTEM

The present invention relates to speech synthesis systems, and in particular to speech systems coding and synthesis systems which can be used in speech communication systems operating at low bit rates.

Speech can be represented as a waveform the detailed structure of which represents the characteristics of the vocal tract and vocal excitation of the person producing the speech. If a speech communication system is to be capable of providing an adequate perceived quality, the transmitted information must be capable of representing that detailed structure. Most of the power in voiced speech is at relatively low frequencies, for example below 2kHz. Accordingly good quality speech synthesis can be achieved on the basis of speech waveforms that have been low pass filtered to reject higher frequency components. The perceived speech quality is however adversely effected if the frequency is restricted much below 4kHz.

Many models have been suggested for defining the characteristics of speech. The known models rely upon dividing a speech signal into blocks or frames and deriving parameters to represent the characteristics of the speech within each frame. Those parameters are then quantized and transmitted to a receiver. At the receiver the quantization process is reversed to recover the parameters, and a speech signal is then synthesised on the basis of the recovered parameters.

The common objective of the designers of the known models is to minimise the volume of data which must be transmitted whilst maximising the perceived quality of the speech that can be synthesised from the transmitted data. In some of the models a distinction is made between whether or not a particular frame is "voiced" or "unvoiced". In the case of voiced speech, speech is produced by glottal excitation and as a result has a quasi-periodic structure. Unvoiced speech is produced by turbulent air flow at a constriction and does not have the "periodic" spectral structure characteristic of voiced speech. Most models seek to take advantage of the fact that voiced speech signals evolve relatively slowly in the context of frames the duration of which is typically 10 to 30msecs. Most models also rely upon quantization schemes intended to minimise the amount of information which must be transmitted without significant loss of perceived quality. As a result of the work done to date it is now possible to produce speech synthesis systems capable of operating at bit rate of only a few thousand bits per second.

One model which has been developed is known as "sinusoidal coding" (R.J. McAulay and T.F. Quatieri, *"Low Rate Speech Coding Based on Sinusoidal Coding"*, Advances in Speech Signal Processing, Editors S. Furui and M.M. Sondhi, Chapter 6, pp. 165-208, Markel Dekker, New York, 1992). This approach relies upon an FFT analysis of each input frame to produce a magnitude spectrum, estimating the pitch period of the input frame from that spectrum, and defining the amplitudes at the pitch related harmonics, the

harmonics being multiples of the fundamental frequency of the frame. An error measure is calculated in the time domain representing the difference between harmonic and aharmonic speech spectra and that error measure is used to define the degree of voicing of the input frame in terms of a frequency value. Thus the parameters used to represent a frame are the pitch period, the magnitude and phase values for each harmonic, and the frequency value. Proposals have been made to operate this system such that phase information is predicted in a coherent way across successive frames.

In another system known as "multiband excitation coding" (D.W. Griffin and J.S. Lim, *"Multiband Excitation Vocoder"* IEEE Transaction on Acoustics, Speech and Signal Processing, vol. 36, pp 1223-1235, 1988 and Digital Voice Systems Inc, *"INMARSAT M Voice Codec, Version 3.0"*, Voice Coding System Description, Module 1, Appendix 1, August 1991) the amplitude and phase functions are determined in a different way from that employed in sinusoidal coding. The emphasis in this system is placed on dividing a spectrum into bands, for example up to twelve bands, and evaluating the voiced/unvoiced nature of each of these bands. Bands that are classified as unvoiced are synthesised using random signals. Where the difference between the pitch estimates of successive frames is relatively small, linear interpolation is used to define the required amplitudes. The phase function is also defined using linear frequency interpolation but in addition includes a constant displacement which is a random variable and which depends on the number of unvoiced bands present in the

short term spectrum of the input signal. The system works in a way to preserve phase continuity between successive frames. When the pitch estimates of successive frames are significantly different, a weighted summation of signals produced from amplitudes and phases derived for successive frames is formed to produced the synthesised signal.

Thus the common ground between the sinusoidal and multiband systems referred to above is that both schemes directly model the input speech signal which is DFT analysed, and both systems are at least partially based on the same fundamental relationship for representing speech to be synthesised. The systems differ however in terms of the way in which amplitudes and phase are estimated and quantized, the way in which different interpolation methods are used to define the necessary phase relationships, and the way in which "randomness" is introduced in the recovered speech.

Various versions of the multiband excitation coding system have been proposed, for example an enhanced multiband excitation speech coder (A. Das and A. Gersho, *Variable-Dimension Spectral Coding of Speech at 2400 bps and below with phonetic classification"*, IEEE Proc. ICASSP-95, pp. 492-495, May 1995) in which input frames are classified into four types, that is noise, unvoiced, fully voiced and mixed voiced, and a variable dimension vector quantization process for spectral magnitude is introduced, the bi-harmonic spectral modelling system (C. Garcia-Matteo., J. L. Alba-Castro and Eduardo R. Banga, *"Speech Coding Using Bi-Harmonic Spectral Modelling"*, Proc. EUSIPCO-94,

Edingburgh, Vol. 2, pp. 391-394, September 1994) in which the short term magnitude spectrum is divided into two bands and a separate pitch frequency is calculated for each band, the spectral excitation coding system (V. Cuperman, P. Lupini and B. Bhattacharya, *"Spectral Excitation Coding of Speech at 2.4 kb/s"*, IEEE Proc. ICASSP-95, pp. 504-507, Detrpot, May 1995) which applies sinusoidal based coding in the linear predictive coding (LPC) residual domain where the synthesised residual signal is the summation of pitch harmonic oscillators with appropriate amplitude and phase functions and amplitudes are quantized using a non-square transformation, the band-widened harmonic vocoder (G. Yang, G Zanellato and H. Leich, *"Band Widened Harmonic Vocoder at 2 to 4 kbps"*, IEEE Proc. ICASSP-95, pp. 504-507, Detroit, May 1995) in which randomness in the signal is introduced by adding jitter to the amplitude information on a per band basis, pitch synchronous multiband coding (H. Yang, S. N. Koh and P. Sivaprakasapilai, *"Pitch Synchronous Multi-Band (PSMB) Speech Coding"*, IEEE Proc. ICASSP-95, pp. 516-519, Detroit, May 1995) in which a CELP (code-excited linear prediction) based coding scheme is used to encode speech period segments, multi band LPC coding (S. Yeldener, M. Kondoz and G. Evans, *"High Quality Multiband LPC Coding of Speech at 2.4 kbits/s"*, Electronic Letters, pp. 1287-1289, Vol. 27, No 14, 4th July 1991) in which a single amplitude value is allocated to each frame to in effect specify a "flat" residual spectrum, and harmonic and noise coding (M. Nishiguchi and J. Matsumoto, *"Harmonic and Noise Coding of LPC Residuals with Classified Vector*

Quantisation", IEEE Proc. ICASSP-95, pp. 484-487, Detroit, May 1995) with classified vector quantization which operates in the LPC residual domain, an input signal being classified as voiced or unvoiced and being full band modelled.

A further type of coding system exists, that is the prototype interpolation coding system. This relies upon the use of pitch period segments or prototypes which are spaced apart in time and reiteration/interpolation techniques to synthesise the signal between two prototypes. Such a system was described as early as 1971 (J.S. Séverwight, "Interpolation Reiterations Techniques for Efficient Speech Transmission", Ph.D. Thesis, Loughborough University, Department of Electrical Engineering, 1971). More sophisticated systems of the same general class have been described more recently, for example in the paper by W.B. Kleijn, "Continuous Representations in Linear Predictive Coding, Proc. ICASSP-91, pp201-204, May 1991. The same author has published a series of related papers. The system employs 20msecs coding frames which are classified as voiced or unvoiced. Unvoiced frames are effectively CELP coded. Pitch prototype segments are defined in adjacent voiced frames, in the LPC residual signal, in a way which ensures maximum alignment (correlation) of the prototypes and defines the prototype so that the main pitch excitation pulse is not near to either of the ends of the prototype. A pitch period in a given frame is considered to be a cycle of an artificial periodic signal from which the prototype for the frame is obtained. The prototypes which have been appropriately

selected from adjacent frames are Fourier transformed and the resulting coefficients are coded using a differential vector quantization scheme.

With this scheme, during synthesis of voiced frames, the decoded prototype Fourier representations for adjacent frames are used to reconstruct the missing signal waveform between the two prototype segments using linear interpolation. Thus the residual signal is obtained which is then presented to an LPC synthesis filter the output of which provides the synthesised voiced speech signal. An amount of randomness can be introduced into voiced speech by injecting noise at frequencies larger than 2khz, the amplitude of the noise increasing with frequency. In addition, the periodicity of synthesised voiced speech is controlled during the quantization of prototype parameters in accordance with a long term signal to change ratio measure that reflects the similarity which exists between the prototypes of adjacent frames in the residual excitation signal.

The known prototype interpolation coding systems rely upon a Fourier Series synthesis equation which involves a linear-with-time-interpolation process. The assumption is that the pitch estimates for successive frames are linearly interpolated to provide a pitch function and an associated instant fundamental frequency. The instant phase used in the cosine and sine terms of the Fourier series synthesis equation is the integral of the instantaneous harmonic frequencies. This synthesis arrangement allows for the linear evolution of the

instantaneous pitch and the non-linear evolution of the instantaneous harmonic frequencies.

A development of this system is described by W.B. Kleijn and J. Haaden, "A Speech Coder Based on Decomposition of Characteristics Waveforms", Proc. ICASSP-95, pp508-511, Detroit, May 1995. In the described system the Fourier series coefficients are low pass filtered over time, with a cut-off frequency of 20Hz, to provide a "slowly evolving" waveform component for the LPC excitation signal. The difference between this low pass component and the original parameters provides the "rapidly evolving" components of the excitation signal. Periodic voice excitation signals are mainly represented by the "slowly evolving" component, whereas random unvoiced excitation signals are represented by the "rapidly evolving" component in this dual decomposition of the Fourier series coefficients. This removes effectively the need for treating voiced and unvoiced frames separately. Furthermore, the rate of quantization and transmission of the two components is different. The "slowly evolving" signal is sampled at relatively long intervals of 25msecs, but the parameters are quantized quite accurately on the basis of spectral magnitude information. In contrast, the spectral magnitude of the "rapidly evolving" signal is sampled frequently, every 4msecs, but is quantized less accurately. Phase information is randomised every 2msecs.

Other developments of the prototype interpolation coding system have been proposed. For example one known system operates on 5msec frames, a

pitch period being selected for voiced frames and DFT transformed to yield prototype spectral magnitude values. These values are quantized and the quantized values for adjacent frames are linearly interpolated. Phase information is defined in a manner which does not satisfy any frequency restrictions at the interpolation boundaries. This causes problems of discontinuity at frame boundaries. At the receiver the excitation signal is synthesised using a decoded magnitude and estimated phase values, via an inverse DFT process. The resulting signal is filtered by a following LPC synthesis filter. This model is purely periodic during voiced speech, and this is why a very short duration frame is used. Unvoiced speech is CELP coded.

The wide range of speech synthesis models currently being proposed, only some of which are described above, and the range of alternative approaches proposed to implement those models, indicates the interest in such systems and the lack of any consensus as to which system provides the most advantageous performance.

It is an object of the present invention to provide an improved low bit rate speech synthesis system.

In known systems in which it is necessary to obtain an estimate of the pitch of a frame of a speech signal, it has been thought necessary, if high quality of synthesised speech is to be achieved, to obtain high resolution non-integer pitch period estimates. This requires complex processes, and it would be highly

desirable to reduce the complexity of the pitch estimation process in a manner which did not result in degraded quality.

According to a first aspect of the present invention, there is provided a speech synthesis system in which a speech signal is divided into a series of frames, and each frame is converted into a coded signal including a voiced/unvoiced classification and a pitch estimate, wherein a low pass filtered speech segment centred about a reference sample is defined in each frame, a correlation value is calculated for each of a series of candidate pitch estimates as the maximum of multiple crosscorrelation values obtained from variable length speech segments centred about the reference sample, the correlation values are used to form a correlation function defining peaks, and the locations of the peaks are determined and used to define a pitch estimate.

The result of the above system is that an integer pitch period value is obtained. The system avoids undue complexity and may be readily implemented.

Preferably the pitch estimate is defined using an iterative process. A single reference sample may be used, for example centred with respect to the respective frame, or alternatively multiple pitch estimates may be derived for each frame using different reference samples, the multiple pitch estimates being combined to define a combined pitch estimate for the frame. The pitch estimate may be modified by reference to a voiced/unvoiced status and/or pitch estimates of adjacent frames to define a final pitch estimate.

The correlation function may be clipped using a threshold value, remaining peaks being rejected if they are adjacent to larger peaks. Peaks are initially selected and can be rejected if they are smaller than a following peak by more than a predetermined factor, for example smaller than 0.9 times the following peak.

Preferably the pitch estimation procedure is based on a least squares error algorithm. Preferably the algorithm defines the pitch as a number whose multiples best fit the correlation function peak locations. Initial possible pitch values may be limited to integral numbers which are not consecutive, the increment between two successive numbers being proportional to a constant multiplied by the lower of those two numbers.

It is well known from the prior art to classify individual frames as voiced or unvoiced and to process those frames in accordance with that classification. Unfortunately such a simple classification process does not accurately reflect the true characteristics of speech. It is often the case that individual frames are made up of both periodic (voiced) and aperiodic (unvoiced) components. Prior attempts to address this problem have not proved particularly effective.

It is an object of the present invention to provide an improved voiced or unvoiced classification system.

According to a second aspect of the present invention there is provided a speech synthesis system in which a speech signal is divided into a series of frames, and each frame is converted into a coded signal including pitch segment

magnitude spectral information, a voiced/unvoiced classification, and a mixed voiced classification which classifies harmonics in the magnitude spectrum of voiced frames as strongly voiced or weakly voiced, wherein a series of samples centred on the middle of the frame are windowed to form a data array which is Fourier transformed to produce a magnitude spectrum, a threshold value is calculated and used to clip the magnitude spectrum, the clipped data is searched to define peaks, the locations of peaks are determined, constraints are applied to define dominant peaks, and harmonics not associated with a dominant peak are classified as weakly voiced.

Peaks may be located using a second order polynomial. The samples may be Hamming windowed. The threshold value may be calculated by identifying the maximum and minimum magnitude spectrum values and defining the threshold as a constant multiplied by the difference between the maximum and minimum values. Peaks may be defined as those values which are greater than the two adjacent values. A peak may be rejected from consideration if neighbouring peaks are of a similar magnitude, e.g. more than 80% of the magnitude, or if there are spectral magnitudes in the same range of greater magnitudes. A harmonic may be considered as not being associated with a dominant peak if the difference between two adjacent peaks is greater than a predetermined threshold value.

The spectrum may be divided into bands of fixed width and a strongly/weakly voiced classification assigned for each band. Alternatively the

frequency range may be divided into two or more bands of variable width, adjacent bands being separated at a frequency selected by reference to the strongly/weakly voiced classification of harmonics.

Thus, the spectrum may be divided into fixed bands, for example fixed bands each of 500Hz, or variable width bands selected in dependence upon the strongly/weakly voiced status of harmonic components of the excitation signal. A strongly/weakly voiced classification is then assigned to each band. The lowest frequency band, e.g. 0-500Hz, may always be regarded as strongly voiced, whereas the highest frequency band, for example 3500Hz to 4000Hz, may always be regarded as weakly voiced, In the event that a current frame is voiced, and the previous frame is unvoiced, other bands within the current frame, e.g. 3000Hz to 3500Hz may be automatically classified as weakly voiced. Generally the strongly/weakly voiced classification may be determined using a majority decision rule on the strongly/weakly voiced classification of those harmonics which fall within the band in question. If there is no majority, alternate bands may be alternately assigned strongly voiced and weakly voiced classifications.

Given the classification of a voiced frame such that harmonics are classified as either strongly or weakly voiced, it is necessary to generate an excitation signal to recover the speech signal which takes into account this classification. It is an object of the present invention to provide such a system.

According to a third aspect of the present invention, there is provided a speech synthesis system in which a speech signal is divided into a series of

frames, each frame is defined as voiced or unvoiced, each frame is converted into a coded signal including a pitch period value, a frame voiced/unvoiced classification and, for each voiced frame, a mixed voiced spectral band classification which classifies harmonics within spectral bands as either strongly or weakly voiced, and the speech signal is reconstructed by generating an excitation signal in respect of each frame and applying the excitation signal to a filter, wherein for each weakly voiced spectral band, an excitation signal is generated which includes a random component in the form of a function which is dependent upon the respective pitch period value.

Thus for each frame which has a spectral band that is classified as weakly voiced, the excitation signal is represented by a function which includes a first harmonic frequency component, the frequency of which is dependant upon the pitch period value appropriate to that frame, and a second random component which is superimposed upon the first component.

The random component may be introduced by reducing the amplitude of harmonic oscillators assigned the weakly voiced classification, for example by reducing the power of the harmonics by 50%, while disturbing the oscillator frequencies, for example by shifting the oscillators randomly in frequency in the range of 0 to 30 Hz such that the frequency is no longer a multiple of the fundamental frequency, and then adding further random signals. The phase of the oscillators producing random signals may be randomised at pitch intervals.

Thus for a weakly voiced band, some periodicity remains but the power of the periodic component is reduced and then combined with a random component.

In a speech synthesis system in which a speech signal is represented in part by spectral information in the form of harmonic magnitude values, it is possible to process an input speech signal to produce a series of spectral magnitude values and then to use all of those magnitude values at harmonic locations in subsequent processing steps. In many circumstances however at least some of the magnitude values contain little information which is useful in the recovery of the input speech signal. Accordingly when magnitude values are quantized for transmission to a receiver it is sensible to discard magnitude values which contain little useful information.

In one known system an input speech signal is processed to produce an LPC residual signal which in turn is processed to provide harmonic magnitude values, but only a fixed number of those magnitude values is vector quantized for transmission to a receiver. The discarded magnitude values are represented at the receiver as identical constant values. This known system reduces redundancy but is inflexible in that the locations of the fixed number of magnitude values to be quantized are always the same and predetermined on the basis of assumption that may be inappropriate in particular circumstances.

It is an object of the present invention to provide an improved magnitude value quantization system.

According to a fourth aspect of the present invention, there is provided a speech synthesis system in which a speech signal is divided into a series of frames, and each voiced frame is converted into a coded signal including a pitch period value, LPC coefficients, and pitch segment spectral magnitude information, wherein the spectral magnitude information is quantized by sampling the LPC short term magnitude spectrum at harmonic frequencies, the locations of the largest spectral samples are determined to identify which of the magnitudes are relatively more important for accurate quantization, and the magnitudes so identified are selected and vector quantized.

Thus rather than relying upon a simple location selection strategy of a fixed number of magnitude values for quantization and transmission, for example the "low part" of the magnitude spectrum, the invention selects only those values which make a significant contribution according to the subjectively important LPC magnitude spectrum, thereby reducing redundancy without compromising quality.

In one arrangement in accordance with the invention a pitch segment of P_n LPC residual samples is obtained, where P_n is the pitch period value of the nth frame, the pitch segment is DFT transformed, the mean value of the resultant spectral magnitudes is calculated, the mean value is quantized and used as a normalisation factor for the selected magnitudes, and the resulting normalised amplitudes are quantized.

Alternatively, the RMS value of the pitch segment is calculated, the RMS value is quantized and used as a normalisation factor for the selected magnitudes, and the resulting normalised amplitudes are quantized.

At the receiver, the selected magnitudes are recovered, and each of the other magnitude values is reproduced as a constant value.

Interpolation coding systems which employ a pitch-related synthesis formula to recover speech generally encounter the problem of coding a variable length, pitch dependant spectral amplitude vector. The quantization scheme referred to above in which only the magnitudes of relatively greater importance are quantized avoids this problem by quantizing only a fixed number of magnitude values and setting the rest of the magnitude values to a constant value. Thus at the receiver a fixed length vector can be recovered. Such a solution to the problem however may result in a relatively spectrally flat excitation model which has limitations in providing high recovered speech quality.

In an ideal world output speech quality would be maximised by quantizing the entire shape of the magnitude spectrum, and various approaches have been proposed for coding the entire magnitude spectrum. In one approach, the spectrum is DFT transformed and coded differentially across successive spectra. This and similar coding schemes are rather inefficient however and operate with relatively high bit rates. The introduction of vector quantization

allowed for the development of sinusoidal and prototype interpolation systems which operate at lower bit rates, typically around 2.4Kbits/sec.

Two vector quantization methodologies have been reported which quantize a variable size input vector with a fixed size code vector. In a first approach, the input vector is transformed to a fixed size vector which is then conventionally vector quantized. An inverse transform of the quantized fixed size vector yields the recovered quantized vector. Transformation techniques which have been used include linear interpolation, band limited interpolation, all pole modelling and non-square transformation. This approach however produces an overall distortion which is the summation of the vector quantization noise and a component which is introduced by the transformation process. In a second known approach, a variable input vector is directly quantized with a fixed size code vector. This approach is based on selecting only a limited number of elements from each codebook vector to form a distortion measure between a codebook vector and an input vector. Such a quantization approach avoids the transformation distortion of the alternative technique mentioned above and results in an overall distortion that is equal to the vector quantization noise, but this is significant.

It is an object of the present invention to provide an improved variable sized spectral vector quantization scheme.

According to a fifth aspect of the present invention, there is provided a speech synthesis system in which a variable size input vector of coefficients to be

transmitted to a receiver for the reconstruction of a speech signal is vector quantized using a codebook defined by vectors of fixed size, the codebook vectors of fixed size are obtained from variable size training vectors and an interpolation technique which is an integral part of the codebook generation process, codebook vectors are compared to the variable sized input vector using the interpolation process, and an index associated with the codebook entry with the smallest difference from the comparison is transmitted, the index being used to address a further codebook at the receiver and thereby derive an associated fixed size codebook vector, and the interpolation process being used to recover from the derived fixed sized codebook vector an approximation of the variable sized input vector.

The invention is applicable in particular to pitch synchronous low bit rate coders of the type described in this document and takes advantage of the underlying principle of such coders which means that the shape of the magnitude spectrum is represented by a relatively small number of equally spaced samples.

Preferably the interpolation process is linear. For an input vector of given dimension, the interpolation process is applied to produce from the codebook vectors a set of vectors of that given dimension. A distortion measure is then derived to compare the interpolated set of vectors and the input vector and the codebook vector which yields the minimum distortion is selected.

Preferably the dimension of the input vectors is reduced by taking into account only the harmonic amplitudes with the input brandwidth range, for

example 0 to 3.4kHz. Preferably the remaining amplitudes i.e. in the region of 3.4kHz to 4 kHz are set to a constant value. Preferably, the constant value is equal to the mean value of the quantized amplitudes.

Amplitude vectors obtained from adjacent residual frames exhibit significant amounts of redundancy which can be removed by means of backward prediction. The backward prediction may be performed on a harmonic basis such that the amplitude value of each harmonic of one frame is predicted from the amplitude value of the same harmonic in the previous frame or frames. A fixed linear predictor may be incorporated in the system, together with mean removal and gain shape quantization processes which operate on a resulting error magnitude vector.

Although the above described variable sized vector quantization scheme provides advantageous characteristics, and in particular provides for good perceived signal quality at a bit rate of for example 2.4Kbits/sec, in some environments a lower bit rate would be highly desirable even at the loss of some quality. It would be possible for example to rely upon a single value representation and quantization strategy on the assumption that the magnitude spectrum of the pitch segment in the residual domain has an approximately flat shape. Unfortunately systems based on this assumption have a rather poor decoded speech quality.

It is an object of the present invention to overcome the above limitation in lower bit rate systems.

According to a sixth aspect of the present invention, there is provided a speech synthesis system in which a speech signal is divided into a series of frames, each frame is converted into a coded signal including an estimated pitch period, an estimate of the energy of a speech segment the duration of which is a function of the estimated pitch period, and LPC filter coefficients defining an LPC spectral envelope, and a speech signal of related power to the power of the input speech signal is reconstructed by generating an excitation signal using spectral amplitudes which are defined from a modified LPC spectral envelope sampled at the harmonic frequencies defined by the pitch period.

Thus, although a single value is used to represent the spectral envelope of the excitation signal, the excitation spectral envelope is shaped according to the LPC spectral envelope. The result is a system which is capable of delivering high quality speech at 1.5Kbits/sec. The invention is based on the observation that some of the speech spectrum resonance and anti-resonance information is also present in the residual magnitude spectrum, since LPC inverse filtering cannot produce a residual signal of absolutely flat magnitude spectrum. As a consequence, the LPC residual signal is itself highly intelligible.

The magnitude values may be obtained by spectrally sampling a modified LPC synthesis filter characteristic at the harmonic locations related to the pitch period. The modified LPC synthesis filter may have reduced feed back gain and a frequency response which consists of equalised resonant peaks, the locations of which are close to the LPC synthesis resonant locations. The value of the feed

back gain may be controlled by the performance of the LPC model such that it is for example proportional to the normalised LPC prediction error. The energy of the reproduced speech signal may be equal to the energy of the original speech waveform.

It is well known that in prototype interpolation coding speech synthesis systems there are often substantial similarities between the prototypes of adjacent frames in the residual excitation signals. This has been used in various systems to improve perceived speech quality by ensuring that there is a smooth evolution of the speech signal over time.

It is an object of the present invention to provide an improved speech synthesis system in which the excitation and vocal tract dynamics are substantially preserved in the recovered speech signal.

According to a seventh aspect of the present invention, there is provided a speech synthesis system in which a speech signal is divided into a series of frames, each frame is converted into a coded signal including LPC filter coefficients and at least one parameter associated with a pitch segment magnitude, and the speech signal is reconstructed by generating two excitation signals in respect of each frame, each pair of excitation signals comprising a first excitation signal generated on the basis of the pitch segment magnitude parameter or parameters of one frame and a second excitation signal generated on the basis of the pitch segment magnitude parameter or parameters of a second frame which follows and is adjacent to the said one frame, applying the

first excitation signal to a first LPC filter the characteristics of which are determined by the LPC filter coefficients of the said one frame and applying the second excitation signal to a second LPC filter the characteristics of which are determined by the LPC filter coefficients of the said second frame, and weighting and combining the outputs of the first and second LPC filters to produce one frame of a synthesised speech signal.

Preferably the first and second excitation signals include the same phase function and different phase contributions from the two LPC filters involved in the above double synthesis process. This reduces the degree of pitch periodicity in the recovered signals. This and the combination of the first and second LPC filter outputs ensures an effective smooth evolution of the speech spectral envelope on a sample by sample basis.

Preferably the outputs of the first and second LPC filters are weighted by half a window function such as a Hamming window such that the magnitude of the output of the first filter is decreasing with time and the magnitude of the output of the second filter is increasing with time.

According to an eighth aspect of the present invention, there is provided a speech coding system which operates on a frame by frame basis, and in which information is transmitted which represents each frame as either voiced or unvoiced and, for each voiced frame, represents that frame by a pitch period value, quantized magnitude spectral information, and LPC filter coefficients, the received pitch period value magnitude spectral information being used to

generate residual signals at the receiver which are applied to LPC speech synthesis filters the characteristics of which are determined by the transmitted filter coefficients, wherein each residual signal is synthesised according to a sinusoidal mixed excitation synthesis process, and a recovered speech signal is derived from the residual signals.

Embodiments of the present invention will now be described, by way of example, with reference to the accompanying drawings, in which:

Figure 1 is a general block diagram of the encoding process in accordance with the present invention;

Figure 2 illustrates the relationship between coding and matrix quantisation frames;

Figure 3 is a general block diagram of the decoding process;

Figure 4 is a block diagram of the excitation synthesis process;

Figure 5 is a schematic diagram of the overlap and add process;

Figure 6 is a schematic diagram of the calculation of an instantaneous scaling factor;

Figure 7 is a block diagram of the overall voiced/unvoiced classification and pitch estimation process;

Figure 8 is a block diagram of the pitch estimation process;

Figure 9 is a schematic diagram of two speech segments which participate in the calculation of a crosscorrelation function value;

Figure 10 is a schematic diagram of speech segments used in the calculation of the crosscorrelation function value;

Figure 11 represents the value allocated to a parameter used in the calculation of the crosscorrelation function value for different delays;

Figure 12 is a block diagram of the process used for calculated the crosscorrelating function and the selection of its peaks;

Figure 13 is a flow chart of a pitch estimation algorithm;

Figure 14 is a flow chart of a procedure used in the pitch estimation process;

Figure 15 is a flow chart of a further procedure used in the pitch estimation process;

Figure 16 is a flow chart of a further procedure used in the pitch estimation process.

Figure 17 is a flow chart of a threshold value selection procedure;

Figure 18 is a flow chart of the voiced/unvoiced classification process;

Figure 19 is a schematic diagram of the voiced/unvoiced classification process with respect to parameters generated during the pitch estimation process;

Figure 20 is a flow chart of the procedure used to determine offset values;

Figure 21 is a flow chart of the pitch estimation algorithm;

Figure 22 is a flow chart of a procedure used to impose constraints on output pitch estimates to ensure smooth evolution of pitch values with time;

Figures 23, 24 and 25 represent different portions of a flow chart of a pitch post processing procedure;

Figure 26 is a general block diagram of the LPC analysis and LPC quantisation process;

Figure 27 is a general flow chart of a strongly or weakly voiced classification process;

Figure 28 is a flow chart of the procedure responsible for the strongly/weakly voiced classification.

Figure 29 represents a speech waveform obtained from a particular speech utterance;

Figure 30 shows frequency tracks obtained for the speech utterance of Figure 29;

Figure 31 shows to a larger scale a portion of Figure 30 and represents the difference between strongly and weakly voiced classifications;

Figure 32 shows a magnitude spectrum of a particular speech segment and the corresponding LPC spectral envelope and the normalised short term magnitude spectra of the corresponding residual segment, excitation segment obtained using a binary excitation model and an excitation segment obtained using the strongly/weakly voiced model;

Figure 33 is a general block diagram of a system for representing and quantising magnitude information;

Figure 34 is a block diagram of an adaptive quantiser shown in Figure 33;

Figure 35 is a general block diagram of a quantisation process;

Figure 36 is a general block diagram of a differential variable size spectral vector quantiser; and

Figure 37 represents the hierarchical structure of a mean gain shape quantiser.

A system in accordance with the present invention is described below, firstly in general terms and then in greater detail. The system operates on an LPC residual signal on a frame by frame basis.

Speech is synthesised using the following general expression:

$$s(i) = \sum_{k=0}^{K} A_k(i)\cos(\vartheta_k(i) + \phi_k) \qquad (1)$$

where i is the sampling instant and $A_k(i)$ represents the amplitude value of the kth cosine term $\cos(\Theta_k(i))$ (with $\Theta_k(i) = \vartheta_k(i) + \phi_k$) as a function of i. In voiced speech K depends on the pitch frequency of the signal.

A voiced/unvoiced classification process allows the coding of voiced and unvoiced frames to be handled in different ways. Unvoiced frames are modelled in terms of an RMS value and a random time series. In voiced frames a pitch period estimate is obtained and used to define a pitch segment which is centred at the middle of the frame. Pitch segments from adjacent frames are DFT transformed and only the resulting pitch segment magnitude information is coded and transmitted. Furthermore, pitch segment magnitude samples are classified as strongly or weakly voiced. Thus in addition to voiced/unvoiced information, the system transmits for every voiced frame the pitch period value, the magnitude spectral information of the pitch segment, the strong/weak voiced classification of the pitch magnitude spectral values, and the LPC coefficient. Thus, the information which is transmitted for every voiced frame is, in addition to voiced/unvoiced information, the pitch period value, the magnitude spectral information of the pitch segment, and the LPC filter coefficients.

At the receiver a synthesis process, that includes interpolation, is used to reconstruct the waveform between the middle points of the current (n+1)th and previous nth frames. The basic synthesis equation for the residual signal is:

$$\mathrm{Re}\, s(i) = \sum_{j=0}^{K} \hat{MG}_j \cos\left(phase_j(i)\right) \qquad (2)$$

where $\hat{MG}_j$ are decoded pitch segment magnitude values and $phase_j(i)$ is calculated from the integral of the linearly interpolated instantaneous harmonic frequencies $\omega_j(i)$. K is the largest value of j for which $\omega_j^n(i) \leq \pi$.

In the transitions from unvoiced to voiced, the initial phase for each harmonic is set to zero. Phase continuity is preserved across the boundaries of successive interpolation intervals.

The synthesis process is performed twice however, once using the magnitude spectral values MG_j^{n+1} of the pitch segment derived from the current (n+1)th frame and again using the magnitude values MG_j^n of the pitch segment derived in the previous nth frame. The phase function $phase_j(i)$ in each case remains the same. The resulting residual signals $Res_n(i)$ and $Res_{n+1}(i)$ are used as inputs to corresponding LPC synthesis filters calculated for the nth and (n+1)th speech frames. The two LPC synthesised speech waveforms are then weighted by $W_{n+1}(i)$ and $W_n(i)$ to yield the recovered speech signal.

Thus the overall synthesis process, for successive voiced frames, can be described by:

$$\begin{aligned} S(i) = {} & W_n(i) \sum_{j=0}^{K} H^n\left(\omega_j''(i)\right) MG_j'' \cos\left[phase_j''(i) + \varphi''\left(\omega_j''(i)\right)\right] \\ & + W_{n+1}(i) \sum_{j=0}^{K} H^{n+1}\left(\omega_j''(i)\right) MG_j''^{+1} \cos\left[phase_j''(i) + \varphi''^{+1}\left(\omega_j''(i)\right)\right]. \end{aligned} \qquad (3)$$

where $H^n\left(\omega_j''(i)\right)$ is the frequency response of the nth frame LPC synthesis filter calculated, at the $\omega_j^n(i)$ harmonic frequency function at the ith instant. $\varphi''\left(\omega_j''(i)\right)$ is the associated phase response of this filter. $\omega_j^n(i)$ and $phase_j^n(i)$ are the frequency and phase functions defined for the sampling instants i, with i covering the middle of the nth frame to the middle of the (n+1)th frame segments. K is the largest value of j for which $\omega_j''(i) \leq \pi$.

The above speech synthesis process introduces two "phase dispersion" terms i.e. $\varphi''\left(\omega''_j(i)\right)$ and $\varphi''^{+1}\left(\omega''_j(i)\right)$ which effectively reduce the degree of pitch periodicity in the recovered signal. In addition, this "double synthesis" arrangement followed by an overlap-add process ensures an effective smooth evolution of the speech spectral envelope (LPC) on a sample by sample basis.

The LPC excitation signal is based on a "mixed" excitation model which allows for the appropriate mixing of periodic and random excitation components in voiced frames on a frequency-band basis. This is achieved by operating the system such that the magnitude spectrum of the residual signal is examined, and applying a peak-picking process, near the ω_j resonant frequencies, to detect possible dominant spectral peaks. A peak associated with a frequency ω_j indicates a high degree of voicing (represented by hv_j=1) for that harmonic. The absence of an adjacent spectral peak, on the other hand, indicates a certain degree of randomness (represented by hv_j=0). When hv_j=1 (to indicate "strong" voicing) the contribution of the jth harmonic to the synthesis process is $M\hat{G}_j \cos\left(phase_j(i)\right)$ However, when hv_j=0 (to indicate "weak" voicing) the frequency of the jth harmonic is slightly dithered, its magnitude $M\hat{G}_j$ is reduced to $\left(M\hat{G}_j / \sqrt{2}\right)$ and random cosine terms are added symmetrically alongside the jth harmonic ω_j. The terms "strong" and "weak" are used in this sense below. The number NRS of these random terms is

$$NRS = 2 \times \left\lceil \frac{\omega_0}{4\pi \times (50/fs)} \right\rceil \qquad (4)$$

where $\lceil\ \rceil$ indicates rounding off to the next larger integer value. Furthermore, the NRS random components are spaced at 50 Hz intervals symmetrically about ω_j. ω_j being located in the middle of such a 50 Hz interval. The amplitudes of the NRS random components are set to $\left(M\hat{G}_j / \sqrt{2 \times NRS}\right)$ Their initial phases are selected randomly from the $[-\pi, +\pi]$ region at pitch period intervals.

The hv_j information must be transmitted to be available at the receiver and, in order to reduce the bit rate allocated to hv_j, the bandwidth of the input signal is divided into a number of fixed size bands BD_k and a "strongly" or "weakly" voiced flag Bhv_k is assigned for each band. In a "strongly" voiced band, a highly periodic signal is reproduced. In a "weakly" voiced band, a signal which combines both periodic and aperiodic components is required. These bands are classified as strongly voiced (Bhv_k=1) or weakly voiced (Bhv_k=0) using a majority decision rule approach on the hv_j classification values of the harmonics ω_j contained within each frequency band.

Further restrictions can be imposed on the strongly/weakly voiced profiles resulting from the classification of bands. For example, the first λ bands may always be strongly voiced i.e. hv_j=1 for BD_k with k=1,2,...,λ, and λ being a variable. The remaining spectral bands can be strongly or weakly voiced.

Figure 1 schematically illustrates processes operated by the system encoder. These processes are referred to in Figure 1 as Processes I to VII and these terms are used throughout this document. Figure 2 represents the relationship between analysis/coding frame sizes employed. These are M samples per coding frame, e.g. 160 samples per frame, and k frames are analysed in a block, for example k=4. This block size is used for matrix quantization. A speech signal is input and processes I, III, IV, VI AND VII produce outputs for transmission.

Assuming that the first Matrix Quantization analysis frame (MQA) of k×M samples is available, each of the k coding frames within the MQA is classified as voiced or unvoiced (V_n) using, Process I. A pitch estimation part of Process I provides a pitch period value P_n only when a coding frame is voiced.

Process II operates in parallel on the input speech samples and estimates p LPC filter coefficients $\underline{a}$ (for example p=10) every L samples (L is a multiple of M i.e. L=m×M, and m may be equal to for example 2). In addition, k/m is an integer and represents the frame dimension of the matrix quantizer employed in Process III. Thus the LPC filter coefficients are quantized, using Process III and transmitted. The quantized coefficients $\underline{\hat{a}}$ are used to derive a residual signal $R^n(i)$.

When an input coding frame is unvoiced, the Energy E_n of the residual obtained for this frame is calculated (Process VII). $\sqrt{E_n}$ is then quantized and transmitted.

When the nth coding frame is classified as voiced, a segment of P_n residual samples is obtained (P_n is the pitch period value associated with the nth frame). This segment is centred in the middle of the frame. The selected P_n samples are DFT transformed (Process V) to yield $\lceil (P_n + 1)/2 \rceil$ spectral magnitude values $M\hat{G}_j''$, $0 \le j < \lceil (P_n + 1)/2 \rceil$, and $\lceil (P_n + 1)/2 \rceil$ phase values. The phase information is neglected. The magnitude information is coded (using Process VI) and transmitted. In addition a segment of 20 msecs, which is centred in the middle of the nth coding frame, is obtained from the residual signal $R^n(i)$. This is input to Process IV, together with P_n to provide the strongly/weakly voiced classification parameters hv_j^n of the harmonics ω_j^n. Process IV produces quantized Bhv information, which for voiced frames is multiplexed and transmitted to the receiver together with the voiced/unvoiced decision V_n, the pitch period P_n, the quantized LPC coefficients $\underline{\hat{a}}$ of the corresponding LPC frame, and the magnitude values $M\hat{G}_j''$. In unvoiced frames only the $\sqrt{E_n}$ quantized value and the quantized LPC filter coefficients $\underline{\hat{a}}$ are transmitted.

Figure 3 schematically illustrates processes operated by the system decoder. In general terms, given the received parameters of the nth coding frame and those of the previous (n-1)th coding frame, the decoder synthesises a speech signal $S_n(i)$ that extends from the middle of

the (n-1)th frame to the middle of the nth frame. This synthesis process involves the generation in parallel of two excitation signals $Res_n(i)$ and $Res_{n-1}(i)$ which are used to drive two independent LPC synthesis filters $1/A_n(z)$ and $1/A_{n-1}(z)$ the coefficients of which are derived from the transmitted quantized coefficients $\hat{\underline{a}}$. The outputs $X_n(i)$ and $X_{n-1}(i)$ of these synthesis filters are weighted and added to provide a speech segment which is then post filtered to yield the recovered speech $S_n(i)$. The excitation synthesis process used in both paths of Figure 3 is shown in more detail in Figure 4.

The process commences by considering the voiced/unvoiced status V_k, where k is equal to n or n-1, (see Figure 4). When the frame is unvoiced i.e. $V_k=0$, a gaussian random number generator RG(0,1) of zero mean and unit variance, provides a time series which is subsequently scaled by the $\sqrt{\hat{E}_k}$ value received for this frame. This is effectively the required:

$$Res_k(i) = \sqrt{\hat{E}_k} \times RG(0,1) \qquad (5)$$

signal which is then presented to the corresponding LPC synthesis filter $1/A_k(z)$, k=n or n-1. Performance could be increased if the $\sqrt{E_k}$ value was calculated, quantized and transmitted every 5msecs. Thus, provided that bits are available when coding unvoiced speech, four $\sqrt{E_{k,\xi}}$, ξ=0,...,3, values are transmitted for every unvoiced frame of 20msecs duration (160 samples).

In the case where $V_k=1$, the $Res_k(i)$ excitation signal is defined as the summation of a "harmonic" $Res_k^h(i)$ component and a "random" $Res_k^r(i)$ component. The top path of the $V_k=1$ part of the synthesis in Figure 4, which provides the harmonic component of this mixed excitation model, calculates always the instantaneous harmonic frequency function $\omega_j^n(i)$ which is associated with the interpolation interval that is defined between the middle points of the nth and (n-1)th frames. (i.e. this action is independent of the value of k). Thus, when

decoding the nth frame, $\omega_j^n(i)$ is calculated using the pitch frequencies $f_j^{1,n}$, $f_j^{2,n}$ and linear interpolation i.e.

$$\omega_j''(i) = 2\pi \frac{f_j^{1,n} - f_j^{2,n}}{M} i + 2\pi f_j^{2,n} \quad (6)$$

with $0 \le j < \lceil (P_{max} + 1)/2 \rceil$, $0 \le i < M$ and $P_{max} = \max[P_n, P_{n-1}]$

The frequencies, $f_j^{1,n}$ and $f_j^{2,n}$ are defined as follows:

I) When both the nth and (n-1)th coding frames are voiced i.e $V_n=1$ and $V_{n-1}=1$, then the pitch frequencies are estimated as follows:

a) If $|P_n - P_{n-1}| \le 0.2 \times (P_n + P_{n-1})$ (7)

which means that the pitch values of the nth and (n-1)th coding frames are rather similar, then:

$$f_j^{1,n} = j\frac{1}{P_n} + (1 - hv_j'') \times RU(-a,+a) \quad (8)$$

$$f_j^{2,n} = \begin{cases} f_j^{1,n-1} + jb & if\ (V_{n-1} = V_{n-2} = 1)\ AND\ (|P_{n-1} - P_{n-2}| > 0.2(P_{n-1} + P_{n-2})) \\ f_j^{1,n-1} & otherwise \end{cases} \quad (9)$$

The $f_j^{1,n-1}$ value is calculated during the decoding process of the previous (n-1)th coding frame. hv_j^n is the strongly/weakly voiced classification (0, or 1) of the jth harmonic ω_j^n. P_n and P_{n-1} are the received pitch estimates from the n and n-1 frames. RU(-a,+a) indicates the output of a random number generator with uniform pdf within the -a to +a range. (a=0.00375)

b) if $|P_n - P_{n-1}| > 0.2 \times (P_n + P_{n-1})$ (10)

then $f_j^{1,n} = j\left(\frac{1}{P_n} - b\right) + (1 - hv_j'') \times RU(-a,+a)$ (11)

and $f_j^{2,n} = f_j^{1,n-1} + b \times j$

where b is defined as:

$$b = \frac{\left| f_j^{1,n-1} - \frac{1}{P_n} \right| - \frac{0.2(P_n + P_{n-1})}{P_n P_{n-1}}}{2} \times \mathrm{sgn}\left(\frac{1}{P_n} - f_j^{1,n-1}\right) \quad (12)$$

Notice that in case (b) which applies for significantly different P_n and P_{n-1} pitch estimates, equations 11 and 12 ensure that the rate of change of the $\omega_j^n(i)$ function is restricted to $\left(j\frac{0.2(P_n+P_{n-1})}{P_nP_{n-1}}\right)\Big/M$.

II) When one of the two coding frames (i.e. n, n-1) is unvoiced, one of the following two definitions is applicable:

a) for $V_{n-1}=0$ and $V_n=1$

$$f_j^{2,n}=\frac{1}{P_n}j \qquad 0\le j<\left\lceil\frac{P_n+1}{2}\right\rceil$$

and $f_j^{1,n}$ is given by Equation (8).

b) for $V_{n-1}=1$ and $V_n=0$

$f_j^{2,n}$ is set to the $fj^{1,n-1}$ value, which has been calculated during the decoding process of the previous (n-1)th coding frame and $f_j^{1,n}=f_j^{2,n}$.

Given $\omega_j^n(i)$ the instantaneous function $phase_j^n(i)$ is calculated by:

$$phase_j^n=2\pi\frac{\left(f_j^{1,n}-f_j^{2,n}\right)}{2M}i^2+2\pi f_j^{2,n}i+phase_j^{n-1}(M) \qquad for\ 0\le j<\left\lceil\frac{P_{\max}+1}{2}\right\rceil \quad and\ 0\le i<M \qquad (13)$$

Furthermore, the "harmonic" component $\mathrm{Re}\,s_k''(i)$ of the residual signal is given by:

$$\mathrm{Re}\,s_k''(i)=\sum_{j=0}^{\left\lceil\frac{P_{\max}+1}{2}\right\rceil-1}C_j(i)\times M\tilde{G}_j^k(hv_j^k)\times\cos\left[phase_j^n(i)\right] \qquad 0\le i<M \qquad (14)$$

where k=n or n-1,

$$C_j(i)=\begin{cases}0 & if\ \omega_j''(i)>\pi\\ 1 & if\ \omega_j''(i)\le\pi\end{cases}$$

$$M\tilde{G}_j^k(hv_j^k)=\begin{cases}\left(M\hat{G}_j^k\right)\Big/\left(\sqrt{2}\right) & for\ hv_j^k=0\\ M\hat{G}_j^k & for\ hv_j^k=1\\ 0 & otherwise,\ including\ j=0\end{cases} \quad and \quad 1\le j<\left\lceil\frac{P_k+1}{2}\right\rceil$$

and

$\hat{MG}_j^k \quad j = 0,\ldots,\lfloor (P_k + 1)/2 \rfloor - 1$ are the received magnitude values of the "kth" coding frame, with k=n or k=n-1.

The second path of the V_k=1 case in Figure 4 provides the random excitation component $\mathrm{Re}\, s_k'(i)$. In particular, given the recovered strongly/weakly voiced classification values hv_j^k, the system calculates for those harmonics with hv_j^k=0 the number of random sinusoidal NRS components, which are used to randomise the corresponding harmonic. This is:

$$NRS = 2 \times \left\lceil \frac{\omega_0^k}{4\pi \times (50/fs)} \right\rceil \qquad (15)$$

where fs is the sampling frequency. Notice that the NRS random sinusoidal components are located symmetrically about the corresponding harmonic ω_j^k and they are spaced 50 Hz apart.

The instantaneous frequency of the qth random component, q=0,1,...,NRS-1, for the jth harmonic ω_j^k is calculated by:

$$\omega_{j,q}^k(i) = \omega_j^k(i) + 2\pi \times (25/fs) + (q - (NRS/2)) \times 2\pi \times (50/fs \quad for \quad 0 < j < \left\lceil \frac{P_{max}+1}{2} \right\rceil \qquad (16)$$

$$and \quad 0 \le i \le M$$

The associated phase value is:

$$Ph_{j,q}^k(i) = \frac{\omega_{j,q}^k(M) - \omega_{j,q}^k(0)}{2M} i^2 + i\omega_{j,q}^k(0) + \varphi_{j,q} \qquad for \quad 0 < j < \left\lceil \frac{P_{max}+1}{2} \right\rceil \qquad (17)$$

$$and \quad 0 \le i \le M$$

where $\varphi_{j,q} = RU(\pi, -\pi)$. In addition, the $Ph_{j,q}^k(i)$ function is randomised at pitch intervals (i.e. when the phase of the fundamental harmonic component is a multiple of 2π, i.e. $\mathrm{mod}(phase_1''(i), 2\pi) = 0$).

Given the $Ph_{j,q}^k(i)$, the random excitation component $Res_{kr}(i)$ is calculated as follows:

$$\mathrm{Re}\, s_k'(i) = \sum_{j=0}^{\left\lceil \frac{P_{max}+1}{2} \right\rceil - 1} \sum_{q=0}^{NRS-1} C_{j,q}(i) \times MG_{j,q}^k(hv_j^k) \times \cos\left(Ph_{j,q}^k(i)\right) \qquad 0 \le i < M \qquad (18)$$

where

$$MG^{k}_{j,q}(hv^{k}_{j}) = \begin{cases} \left(M\hat{G}^{k}_{j}\right) / \left(\sqrt{2NRS}\right) & for\ hv^{k}_{j} = 0 \\ 0 & for\ hv^{k}_{j} = 1 \\ 0 & otherwise, including\ j = 0 \end{cases} \quad and \quad 1 \le j < \left\lceil \frac{P_k + 1}{2} \right\rceil$$

$$C_{j,q}(i) = \begin{cases} 0 & \omega^{k}_{j,q}(i) > \pi \\ 1 & \omega^{k}_{j,q}(i) \le \pi \end{cases}$$

Thus for V_k=1 voiced coding frames, the mixed excitation residual is formed as:

$$\mathrm{Re}\, s_k(i) = \mathrm{Re}\, s^{h}_{k}(i) + \mathrm{Re}\, s^{r}_{k}(i) \qquad \textbf{(19)}$$

Notice that when V_k=0, instead of using Equation 5, the random excitation signal $Res_k(i)$ can be generated by the summation of random cosines located 50 Hz apart, where their phase is randomised every λ samples, and λ<M, i.e

$$\mathrm{Re}\, s_k(i) = \sum_{j=1}^{80} \sqrt{\frac{E_k}{40}} \cos\left(2\pi (fs/50) i + \delta(i - \lambda \times \xi - \zeta) \times RU(-\pi, +\pi)\right) \quad where \qquad \textbf{(20)}$$

$$\xi = 0,1,2,\ldots, and\ 0 \le i < M \qquad and$$

ζ is defined so as to ensure that the phase of the cos terms is randomised every λ samples across frame boundaries. The resulting $Res_n(i)$ and $Res_{n-1}(i)$ excitation sequences, see Figure 4, are processed by the corresponding $1/A_n(z)$ and $1/A_{n-1}(z)$ LPC synthesis filters. When coding the next (n+1)th frame, $1/A_{n-1}(z)$ becomes $1/A_n(z)$ (including the memory) and $1/A_n(z)$ becomes $1/A_{n+1}(z)$ with the memory of $1/A_n(z)$. This is valid in all cases except during an unvoiced to voiced transition, where the memory of the $1/A_{n+1}(z)$ filter is set to zero. The coefficients of the $1/A_n(z)$ and $1/A_{n-1}(z)$ synthesis filters are calculated directly from the nth and (n-1)th coding speech frames respectively, when the LPC analysis frame size L is equal to M samples. However, when L≠M (usually L>M) linear interpolation is used on the filter coefficients (defined every L samples) so that the transfer function of the synthesis filter is updated every M samples.

The output signals of these filters, denoted as $X_{n-1}(i)$ and $X_n(i)$, are weighted, overlapped and added as schematically illustrated in Figure 5 to yield $\hat{X}_n(i)$ i.e:

$$\hat{X}_n(i) = W_{n-1}(i)X_{n-1}(i) + W_n(i)X_n(i)$$

where

$$W_n(i) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi}{2M-1}i\right) & for\ 0 \le i < M \quad when\ V_n = V_{n-1} \\ \begin{cases} 0 & for\ 0 \le i < 0.25M \\ 0.5 - 0.5\cos\left(\pi\dfrac{i-0.25M}{0.5M-1}\right) & for\ 0.25M \le i < 0.75M \\ 1 & for\ 0.75M \le i < M \end{cases} & when\ V_n \ne V_{n-1} \end{cases} \quad (21)$$

and

$$W_{n-1}(i) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi(i+M-0.5)}{2M-1}\right) & for\ 0 \le i < M \quad when\ V_n = V_{n-1} \\ \begin{cases} 1 & for\ 0 \le i < 0.25M \\ 0.5 + 0.5\cos\left(\pi\dfrac{i-0.25M}{0.5M-1}\right) & for\ 0.25M \le i < 0.75M \\ 0 & for\ 0.75M \le i < M \end{cases} & when\ V_n \ne V_{n-1} \end{cases} \quad (22)$$

$\hat{X}_n(i)$ is then filtered via a PF(z) post filter and a high pass filter HP(z) to yield the speech segment $S'_n(i)$. PF(z) is the conventional post filter:

$$PF(z) = \frac{A(z/b)}{A(z/c)}\left(1 - \mu z^{-1}\right) \quad (23)$$

with b=0.5, c=0.8 and $\mu = 0.5K_1''.K_1''$ is the first reflection coefficient of the nth coding frame. HP(z) is defined as:

$$HP(z) = \frac{b^1 - c_1 z^{-1}}{1 - a_1 z^{-1}} \qquad (24)$$

with $b_1=c_1=0.9807$ and $a_1=0.961481$.

In order to ensure that the energy of the recovered S(i) signal is preserved, as compared to that of the $\hat{X}(i)$ sequence, a scaling factor SC is calculated every LPC frame of L samples.

$$SC_l = \sqrt{\frac{E_l'}{\tilde{E}_l}} \qquad (25)$$

where: $E_l' = \sum_{i=0}^{L-1} \hat{X}_l(i)^2$ and $\tilde{E}_l = \sum_{i=0}^{L-1} S_l'(i)^2$

SC_l is associated with the middle of the lth LPC frame as illustrated in Figure 6. The filtered samples from the middle of the (l-1)th frame to the middle of the lth frame are then multiplied by $SC_l(i)$ to yield the final output of the system, $S_l(i)=SC_l(i)\times S'_l(i)$ where:

$$SC_l(i) = SC_l W_l(i) + SC_{l-1} W_{l-1}(i) \qquad 0 \le i < L \qquad (26)$$

and

$$W_l(i) = 0.5 - 0.5\cos\left(\pi \frac{i}{L-1}\right) \qquad 0 \le i < L$$

$$W_{l-1}(i) = 0.5 + 0.5\cos\left(\pi \frac{i}{L-1}\right) \qquad 0 \le i < L$$

The scaling process introduces an extra half LPC frame delay into the coding-decoding process.

The above described energy scaling procedure operates on an LPC frame basis in contrast to both the decoding and PF(z), HP(z) filtering procedures which operate on the basis of a frame of M samples.

Details of the coding processes represented in Figure 1 will now be described.

Process I derives a voiced/unvoiced (V/UV) classification V_n for the nth input coding frame and also assigns a pitch estimate P_n to the middle sample M_n of this frame. This process is illustrated in Figure 7.

The V/UV and pitch estimation analysis frame is centred at the middle M_{n+1} of the (n+1)th coding frame with 237 samples on either side. The signal x(i) in the above analysis frame is low pass filtered with a cut off frequency f_c=1.45KHz and the resulting (-147, 147) samples centred about M_{n+1} are used in a pitch estimation algorithm, which yields an estimate $P_{M_{n+1}}$. The pitch estimation algorithm is illustrated in Figure 8, where P represents the output of the pitch estimation process. The 294 input samples are used to calculate a crosscorrelation function CR(d), where d is shown in Figure 9 and 20≤d≤147. Figure 9 shows the two speech segments which participate in the calculation of the crosscorrelation function value at "d" delay. In particular, for a given value of d, the crosscorrelation function $\rho^d(j)$ is calculated for the segments $\{x_L\}^d$, $\{x_R\}^d$,as:

$$\rho^d(j) = \frac{\sum_{i=0}^{d-j-1} \left(\left(x_L^d(i) - \bar{x}_L^d \right) \left(x_R^d(i) - \bar{x}_R^d \right) \right)}{\sqrt{\sum_{i=0}^{d-j-1} \left(x_L^d(i) - \bar{x}_L^d \right)^2} \sqrt{\sum_{i=0}^{d-j-1} \left(x_R^d(i) - \bar{x}_R^d \right)^2}} \qquad (27)$$

where:

x_L^d (i)=x(M_{n+1}-d+j+i), x_R^d (i)=x(M_{n+1}+j+i), for 0≤i≤d-j-1, j=0,1,...,f(d) (Figure 10 schematically represents the X_L^d *and* X_R^d speech segments used in the calculation of the value $CR(d)$ and the non linear relationship between d and f(d) is given in Figure 11 $\bar{x}_L^d$ and $\bar{x}_R^d$ represent the mean value of the $\{x_L\}^d$ and $\{x_R\}^d$ sequences respectively.

The algorithm then selects max[ρ^d(j)] and defines CR(d)= $\max_{0\le j\le f(d)}$ [ρ^d(j)], 20≤d≤147.

In addition to CR(d), the box in Figure 8 labelled "Calculation of CR function and selection of its peaks", whose detailed diagram is shown in Figure 12, provides also the locations loc(k)

of the peaks of the CR(d) function, where k=1,2,...,Np and Np is the number of peaks in a CR(d) function.

Figure 12 is a block diagram of the process involving the calculation of the CR function and the selection of its peaks. As illustrated, given CR(d), a threshold th(d) is determined as:

$$th(d) = CR\left(d_{\max}^{n+1}\right) - b - \left(d - d_{\max}^{n+1}\right) \times a - c \qquad (28)$$

where c=0.08 when $(V_n' = 1)\,AND\left[\left|d_{\max}'' - P_n'\right| < 0.15 \times \left(d_{\max}'' + P_n'\right)\right] OR(V_{n-1} = 1)]$
$AND(d > 0.875 \times P_n')\,AND(d < 1.125 \times P_n')$

or c=0 elsewhere.

and constants a and b are defined as:

b	0.025	0.04	0.05
a	0.0005	0.0005	0.0006
V'n	1	1	0
V_{n-1}	1	0	1/0

$d_{\max}^{n+1}$ is equal to the value of d for which CR(d) is maximised to $CR_{M_{n-1}}^{\max}$. Using this threshold the CR(d) function is clipped to $CR_L(d)$. i.e.

$CR_L(d) = 0$ for $CR(d) \le th(d)$

$CR_L(d) = CR(d)$ otherwise.

$CR_L(d)$ contains segments G_s s=1,2,3....., of positive values separated by G_0 runs of zero values. The algorithm examines the length of the G_0 runs which exist between successive G_s segments (i.e. G_s and G_{s+1}), and when $G_0 < 17$, then the G_s segment with the max $CR_L(d)$ value is kept. This procedure yields $C\hat{R}_L(d)$, which is then examined by the following "peak picking" procedure. In particular those $C\hat{R}_L(d)$ values are selected for which:

$$C\hat{R}_L(d) > C\hat{R}_L(d-1) \quad \text{and} \quad C\hat{R}_L(d) > C\hat{R}_L(d+1)$$

However certain peaks can be rejected if:

$C\hat{R}_L(loc(k)) \le C\hat{R}_L(loc(k+1)) \times 0.9$

This ensures that the final $C\hat{R}_L(loc(k))$ k=1,...,Np does not contain spurious low level $C\hat{R}_L(d)$ peaks. The locations d of the above defined $C\hat{R}_L(d)$ peaks are given by loc(k) k=1,2,...,Np.

CR(d) and loc(k) are used as inputs to the following Modified High Resolution Pitch Estimation algorithm (MHRPE) shown in Figure 8, whose output is $P_{M_{mt}}$. The flowchart of this MHRPE procedure is shown in Figure 13, where P is initialised with 0 and, at the end, the estimated P is the requested $P_{M_{mt}}$. In Figure 13 the main pitch estimation procedure is based on a Least Squares Error (LSE)-algorithm which is defined as follows:

For each possible pitch value j in the range from 21 to 147 with an increment of $0.1 \times j$. i.e. $j \in \{21,23,25,27,30,33,36,40,44,48,53,58,64,70,77,84,92,101,111,122,134\}$. (Thus 21 iterations are performed.)

1) Form the multiplication factor vector:

$$\bar{u}_j = \left[\frac{1}{j}\overline{loc}\right]$$

2) Reject possible pitch j and go back to *(1)* if

a) the same element occurs in $\bar{u}_j$ twice.

b) the elements of $\bar{u}_j$ have as a common factor a prime number.

3) Form the following error quantity

$$E_j = \overline{loc}^T \overline{loc} - 2p_j \bar{u}_j^T \overline{loc} + p_j^2 \bar{u}_j^T \bar{u}_j$$

where

$$p_j = \frac{\overline{loc}^T \bar{u}_j}{\bar{u}_j \bar{u}_j^T}$$

4) Select the p_{js} value for which the associated Error quantity E_{js} is minimum. (i.e. $js: E_{js} \le E_j \quad \forall j \in \{21,23,...134\}$). Set $P=p_{js}$.

The next two general conditions "Reject Highest Delay" loc(Np) and "Reject Lowest Delay" loc(1) are included in order to reject false pitch, "double" or "half" values and in general to provide constraints in the pitch estimates of the system. The "Reject Highest Delay" condition involves 3 constraits:

i) if P=0 then reject loc(Np).

ii) if loc(Np) >100 then find the local maximum $CR(d_{lm})$ in CR(d) at the vicinity of the estimated pitch P (i.e $0.8\times P$ to $1.2\times P$) and compare this with $th(d_{lm})$, which is determined as in Equation 28 Reject loc(Np) when $CR(d_{lm})<th(d_{lm})-0.02$.

iii) If the error E_{js} of the LSE algorithm is larger than 50 and $\bar{u}_{p_n}(Np)=Np$ with Np>2 then reject loc(Np).

The flowchart of this is given in Figure 14.

The "Reject Lowest Delay" general condition, whose flowchart is given in Figure 15, rejects loc(1) when the following three constraints are simultaneously satisfied:

i) The density of detection of the peaks of the correlation coefficient function is less than or equal to 0.75. i.e.

$$\frac{Np}{u_{p_n}(Np)} \le 0.75$$

ii) If the location of the first peak is neglected (i.e. loc(1)), then the remaining locations exhibit a common factor.

iii) The value of the correlation coefficient function at the locations of the missing peaks is relatively small compared to adjacent detected peaks. i.e.

If $u_{Pn}^{k}-u_{Pn}(k)>1$, for k=1,...Np. then

for $i=u_{Pn}(k)+1 : u_{Pn}(k+1)-1$

a) find local maximum $CR(d_{lm})$ in the range from $(i-0.1)\times loc(1)$ to $(i+0.1)\times loc(1)$.

b) if $CR(d_{lm}) <0.97\times CR(u_{Pn}(k))$ then Reject Lowest Delay, **END**.

else **Continue**

This concludes the pitch estimation procedure of Figure 7 whose output is $P_{M_{n+1}}$. As is also illustrated in Figure 7 however, in parallel to the pitch estimation, Process I obtains 160 samples centred at the middle of the M_{n+1} coding frame, removes their mean value, and then calculates R0, R1 and the average R_{av} of the energies of the previous K non-silence coding frames. K is fixed to 50 for the first 50 non-silence coding frames, increases from 50 to 100 with the next 50 non-silence coding frames, and then remains constant at the value of 100. The flowchart of the procedure that calculates R_{av}, R1, R0 and updates the R_{av} buffer is shown in Figure 16, where "Count" represents the number of non-silence speech frames, and "++" denotes increase by one. Notice that TH is an adaptive threshold that is representative of a silence (non speech) frame and is defined as in Figure 17. CR in this case is equal to $CR_{M_{n+1}}^{max}$.

Given R0, R1, R_{av} and $CR_{M_{n+1}}^{max}$, the V/UV part of Process I calculates the status $V_{M_{n+1}}$ of the n+1 frame. The flowchart of this part of the algorithm is shown in Figure 18 where "V" represents the output V/UV flag of this procedure. Setting the "V" flag to 1 or 0 indicates voiced or unvoiced classification respectively. The "CR" parameter denotes the maximum value of the CR function which is calculated in the pitch estimation process. A diagrammatic representation of the voiced/unvoiced procedure is given in Figure 19.

Having the $V_{M_{n+1}}$ value, the $P_{M_{n+1}}$ estimate and the V'_n and P'_n estimates which have been produced from Process I operating on the previous nth coding frame, as illustrated in Figure 7, part b, two further locations M_{n+1}+d1 and M_{n+1}+d2 are estimated and the corresponding [-147,147] segments of filtered speech samples are obtained as illustrated in Figure 7, part b. These additional two analysis frames are used as input to the "Pitch Estimation process" of

Figure 8 to yield $P_{M_{n+1}+d1}$ and $P_{M_{n+1}+d2}$. The procedure for calculating d1 and d2 is given in the flowchart of Figure 20.

The final step in part (a) of Process I of Figure 7, evolves the previous V/UV classification procedure of Figure 8 with inputs R0, R1, R_{av}, and

$$CR = \max\left[CR^{\max}_{M_{n+1}}, CR^{\max}_{M_{n+1}+d1}, CR^{\max}_{M_{n+1}+d2}\right]$$

to yield a preliminary value V^{pr}_{n+1}.

In addition, a multipoint pitch estimation algorithm accepts $P_{M_{n+1}}$, $P_{M_{n+1}+d1}$, $P_{M_{n+1}+d2}$, V_{n-1}, P_{n-1}, V'_n, P'_n to provide a preliminary pitch value P^{pr}_{n+1}. The flowchart of this multipoint pitch estimation algorithm is given in Figure 21, where P_1, P_2 and P_0 represent the pitch estimates associated with the M_{n+1+d1}, M_{n+1+d2} and M_{n+1} points respectively, and P denotes the output pitch estimate of the process, that is P_{n+1}.

Finally part (b) Process I of Figure 7 imposes constraints on the V^{pr}_{n+1} and P^{pr}_{n+1} estimates in order to ensure a smooth evolution for the pitch parameter. The flowchart of this section is given in Figure 22. At the start of this process "V" and "P" represent the voicing flag and pitch estimate values before constraints are applied (V^{pr}_{n+1} and P^{pr}_{n+1} in Figure 7) whereas at the end of the process "V" and "P" represent the voicing flag and pitch estimate values after the constraints have been applied (V'_{n+1} and P'_{n+1}). The V'_{n+1} and P'_{n+1} produced from this section are then used in the next pitch past processing section together with V_{n-1}, V'_n, P_{n-1} and P'_n to yield the final voiced/unvoiced and pitch estimate parameters V_n and P_n for the nth coding frame. This pitch post processing stage is defined in the flowchart of Figures 23, 24 and 25, the output A of Figure 23 being the input to Figure 24, and the output B of Figure 24 being the input to Figure 25. At the start of this procedure "P_n" and "V_n" represent the pitch estimate and voicing flag respectively, which correspond to the nth coding frame prior to post processing (i.e. P'_n, V'_n) whereas at the end of the procedure "P_n" and "V_n" represent the final pitch estimate and voicing flag associated with the nth frame (i.e. P_n, V_n).

The LPC analysis process (Process II of Figure 1) can be performed using the Autocorrelation, Stabilised Covariance or Lattice methods. The Burg algorithm was used, although simple autocorrelation schemes could be employed without a noticeable effect in the decoded speech quality. The LPC coefficients are then transformed to an LSP representation. Typical values for the number of coefficients are 10 to 12 and a 10th order filter has been used. LPC analysis processes are well known and described in the literature, for example "Digital Processing of Speech Signals", L.R. Rabiner and R.W. Schafer, Prentice - Hall Inc., Englewood Cliffs, New Jersey, 1978. Similarly, LSP representations are well known, for example from "Line Spectrum Pair and Speech Data Compression", F Soong and B.H. Juang, Proc. ICASSP-84, pp 1.10.1-1.10.4, 1984. Accordingly these processes and representations will not be described further in this document.

In process II, ten LSP coefficients are used to represent the data. These 10 coefficients could be quantized using scalar 37 bits with the following bit allocation pattern [3,4,4,4,4,4,4,4,3,3]. This is a relatively simple process, but the resulting bit rate of 1850 bits/second is unnecessarily high. Alternatively the LSP coefficients can be Vector Quantised (VQ) using a Split-VQ technique. In the Split-VQ technique an LSP parameter vector of dimension "p" is split into two or more subvectors of lower dimensions and then, each subvector is Vector Quantised separately (when Vector Quantising the subvectors a direct VQ approach is used). In effect, the LSP transformed coefficient vector, C, which consists of "p" consecutive coefficients $(c_1,c_2,...,c_p)$ is split into "K" vectors, C^k $(1\leq k\leq K)$, with the corresponding dimensions d_k $(1\leq d_k\leq p)$. $p=d_1+d_2+...+d_k$. In particular, when "K" is set to "p" (i.e. when C is partitioned into "p" elements) the Split-VQ becomes equivalent to Scalar Quantisation. On the other hand, when K is set to unity (K=1, $d_k=p$) the Split-VQ becomes equivalent to Full Search VQ.

The above Split VQ approach leads to an LPC filter bit rate of the order of 1.3 to 1.4Kbits/sec. In order to minimize further the bit rate of the voice coded system described in this document a Split Matrix VQ (SMQ) has been developed in the University of Manchester and reported in "Efficient Coding of LSP Parameters using Split Matrix Quantisation", C.Xydeas and C.Papanastasiou, Proc ICASSP-95, pp 740-743, 1995. This method results in transparent LPC quantisation at 900bits/sec and offers a flexible way to obtain, for a given quantisation accuracy, the required memory/complexity characteristics for Process III. An important feature of SMQ is a new weighted Euclidean distance which is defined in detail as follows.

$$D(\underline{L}_k(l), \underline{L}'_k(l)) = \sum_{s=0}^{m(k)-1} \left[\sum_{t=0}^{N-1} (LSP^{l+t}_{S(k-1)+s} - LSP'^{l+t}_{S(k-1)+s})^2 w_s(s,t)^2 w_t(t)^2 \right] \quad (29)$$

where $\underline{L}'_k(l)$ represents the kth (k=1,....K) quantized submatrix and $LSP'^{l+t}_{S(k-1)+s}$ are its elements. m(k) represents the spectral dimension of the kth submatrix and N is the SMQ frame dimension. Note also that : $S(k) = \sum_{j=0}^{k} m(j)$, m(0) = 1 and $\sum_{k=1}^{K} m(k) = p$

$$w_t(t) = \left[(1 - Er(t)) \cdot \frac{En(t)}{Aver(En)} \right]^{\alpha} . E_n(t)^{\alpha 1} \quad for\ transmission\ frames\ \ 0 \le t \le N-1 \quad (30)$$

when the N LPC frames consist of both voiced and unvoiced frames

$w_t(t) = En(t)^{\alpha 1}$ otherwise

where Er(t) is the normalised energy of the prediction error of the (l+t)th frame, En(t) is the RMS value of the (l+t)th speech frame and Aver(En) is the average RMS value of the N LPC frames used in SMQ. The values of the constants α and $\alpha 1$ are set to 0.2 and 0.15 respectively.

Also:

$$w_{s,k}(s,t) = \left| LSP^{s+t}_{s(k-1)+s} \right|^{\beta} \quad (31)$$

where $P(l_{k+s}^{n+l})$ is the value of the power envelope spectrum of the (l+t) speech frame at the l_{k+s} $LSP_{s(k-1)+s}^{n+l}$ frequency. β is equal to 0.15

The overall SMQ quantisation process that yields the quantised LSP coefficients vectors $\hat{\underline{l}}^{l}$ to $\hat{\underline{l}}^{l+N-1}$ for the l to l+N-1 analysis frames is shown in Figure 26. This figure also includes the inverse process, which accepts the above $\hat{\underline{l}}^{l+i}$ vectors i=0,..,N-1 and provides the corresponding LPC coefficients vector $\underline{a}^{l}$ to $\hat{\underline{a}}^{l+N-1}$. The $\underline{a}^{l+i}$ i=0,..,N-1, coefficients vectors are modified, prior to the LPC to LSP transformation, by a 10 Hz bandwidth expansion as indicated in Figure 26. A 5Hz bandwidth expansion is also included in the inverse quantisation process.

Process IV of Figure 1 will now be described. This process is concerned with the mixed voiced classification of harmonics. When the nth coding frame is classified as voiced, the residual signal $R^n(i)$ of length 160 samples centred at the middle M_n of the nth coding frame and the pitch period P_n for that frame are used to determine the strongly voiced (hv_j=1)/weakly voiced (hv_j=0) classification associated with the jth harmonic ω_j^n. The flowchart of Process IV is given in Figure 27. The $\underline{R}^n$ array of 160 samples is Hamming windowed and augmented to form a 512 size array, which is then FFT processed. The maximum and minimum values MGR_{max}, MGR_{min} of the resulting 256 spectral magnitude values are determined, and a threshold TH0 is calculated. TH0 is then used to clip the magnitude spectrum. The clipped MGR array is searched to define peaks MGR(P) satisfying:

$$MGR(P)>MGR(P+1) \text{ and } MGR(P)>MGR(P-1)$$

For each peak, MGR(P), "supported" by the MGR(P+1) and MGR(P-1) values a second order polynomial is fitted and the maximum point of this curve is accepted as MGR(P) with a location loc(MGR(P)). Further constraints are then imposed on these magnitude peaks. In particular peaks are rejected :

a) if there are spectral peaks in the neighbourhood of loc(MGR(P)) (i.e in the range (loc(MGR(P))-fo/2 to loc(MGR(P))+fo/2 where fo is the fundamental frequency in Hz), whose value is larger than 80% of MGR(P) or

b) if there are any spectral magnitudes in the same range whose value is larger than MGR(P).

After applying these two constraints the remaining spectral peaks are characterised as "dominant" peaks. The objective of the remaining part of the process is to examine if there is a "dominant" peak near a given harmonic $j\times\omega_0$, in which case the harmonic is classified as strongly voiced and $hv_j=1$, otherwise $hv_j=0$. In particular, two thresholds are defined as follows:

$$TH1=0.15\times fo,\quad TH2=(1.5/P_n)\times fo$$

with $fo=(1/P_n)\times fs$ and fs is the sampling frequency.

The difference $\left(\mathrm{loc}\left(\mathrm{MGR}_d(k)\right)-\mathrm{loc}\left(\mathrm{MGR}_d(k-1)\right)\right)$ is compared to 1.5×fo+TH2, and if larger a related harmonic is not associated with a "dominant" peak and the corresponding classification hv is zero (weakly voiced). $\left(\mathrm{loc}\left(\mathrm{MGR}_d(k)\right)\right)$ is the location of the kth dominant peak and k=1,...,D where D is the number of dominant peaks. This procedure is described in detail in Figure 28, in which it should be noted that the harmonic index j does not always correspond to the magnitude spectrum peak index k, and loc(k) is the location of the kth dominant peak, i.e. $\mathrm{loc}(\mathrm{MGR}_d(k)) = \mathrm{loc}(K)$.

In order to minimise the bit rate associated with the transmission of the hv_j information, two schemes have been employed which coarsely represent hv.

Scheme I

The spectrum is divided into bands of 500Hz each and a strongly voiced/weakly voiced flag Bhv is assigned for each band. The first and last 500Hz bands i.e. 0 to 500 and 3500 to

4000Hz are always regarded as strongly voiced (Bhv=1) and weakly voiced (Bhv=0) respectively. When V_n=1 and V_{n-1}=1 the 500 to 1000 Hz band is classified as voiced i.e. Bhv=1. Furthermore, when V_n=1 and V_{n-1}=0 the 3000 to 3500 Hz band is classified as weakly voiced i.e. Bhv=0. The Bhv values of the remaining 5 bands are determined using a majority decision rule on the hv_j values of the j harmonics which fall within the band under consideration. When the number of harmonics for a given band is even and no clear majority can be established i.e. the number of harmonics with hv_j=1 is equal to the number of harmonics with hv_j=0, then the value of Bhv for that band is set to the opposite of the value assigned to the immediately preceding band. At the decoding process the hv_j of a specific harmonic j is equal to the Bhv value of the corresponding band. Thus the hv information may be transmitted with 5 bits.

Scheme II

In this case the 680 Hz to 3400 Hz range is represented by only two variable size bands. When V_n=1 and V_{n-1}=0 the Fc frequency that separates these two bands can be one of the following:

(A) 680, 1360, 2040, 2720.

whereas, when V_n=1 and V_{n-1}=1, Fc can be one of the following frequencies:

(B) 1360, 2040, 2720, 3400.

Furthermore, the 0 to 680 and 3400 to 4000 Hz bands are always represented with Bhv=1 and Bhv=0 respectively. The Fc frequency is selected by examining the three bands sequentially defined by the frequencies in (A) or (B) and by using again a majority rule on the harmonics which fall within a band. When a band with a mixed voiced classification Bhv=0 is found, i.e. the number of harmonics with hv_j=0 is larger than to the number of harmonics with hv_j=1, then Fc is set to the lower boundary of this band and the remaining spectral region is classified as Bhv=0. In this case only 2 bits are allocated to define Fc. The lower band is strongly voiced with Bhv=1, whereas the higher band is weakly voiced with Bhv=0.

To illustrate the effect of the mixed voice classification on the speech synthesised from the transmitted information, Figures 29 and 30 represent respectively, an original speech waveform obtained for the utterance "Industrial shares were mostly a" and frequency tracks obtained for that utterance. The horizontal axis represents time in terms of frames each of 20msec duration. Figure 31 shows to a larger scale a section of Figure 30, and represents frequency tracks by full lines for the case when the voiced frames are all deemed to be strongly voiced (hv=1) and by dashed lines when the strongly/weakly voiced classification is taken into account so as to introduce random perturbations when hv=0.

Figure 32 shows four waveforms A, B, C and D. Waveform A represents the magnitude spectrum of a speech segment and the corresponding LPC spectral envelope ($\log_{10}$ domain). Waveforms B, C and D represent the normalised Short-Term magnitude spectrum of the corresponding residual segment (B), the excitation segment obtained using the binary (voiced/unvoiced) excitation model (C), and the excitation segment obtained using the strongly voiced/weakly voiced/unvoiced hybrid excitation model (D). It will be noted that the hybrid model introduces an appropriate amount of randomness where required in the $3\pi/4$ to π range such that curve D is a much closer approximation to curve B than curve C.

Process V of Figure 1 will now be described. Once the residual signal has been derived, a segment of P_n samples is obtained in the residual signal domain. The magnitude spectrum of the segment, which contains excitation source information, is derived by applying a P_n points DFT. An alternative solution, in order to avoid the computational complexity of the P_n points DFT, is to apply a fix length FFT (128 points) and to find the value of the magnitude spectrum at the desired points, using linear interpolation.

For a real-valued sequence x(i) of P points the DFT may be expressed as:

$$X(k) = \sum_{i=0}^{P_n-1} x(i)\cos(\frac{2\pi ki}{P}) - j\sum_{i=0}^{P_n-1} x(i)\sin(\frac{2\pi ki}{P})$$

The P_n point DFT will yield a double-side spectrum. Thus, in order to represent the excitation signal as a superposition of sinusoidal signals, the magnitude of all the non DC components must be multiplied by a factor of 2. The total number of single side magnitude spectrum values, which are used in the reconstruction process, is equal to $\lceil (P_n + 1)/2 \rceil$.

Process VI of Figure 1 will now be described. The DFT (Process V) applied on the P_n samples of a pitch segment in the residual domain, yields $\lceil (P_n + 1)/2 \rceil$ spectral magnitudes (MG_j^n, $0<j<\lceil (P_n + 1)/2 \rceil$) and $\lceil (P_n + 1)/2 \rceil$ phase values. The phase information is neglected. However, the continuity of the phase between adjacent voiced frames is preserved. Moreover, the contribution of the DC magnitude component is assumed to be negligible and thus, MG_0^n is set to 0. In this way, the non-DC magnitude spectrum is assumed to contain all the perceptually important information.

Based on the assumption of an "approximately" flat shape magnitude spectrum for the pitch residual segment, various methods could be used to represent the entire magnitude spectrum with a single value. Specifically, a modified single value spectral amplitude representation (MSVSAR) technique is described below.

MSVSAR is based on the observation that some of the speech spectrum resonance and anti-resonance information is also present at the residual magnitude spectrum (G.S. Kang and S.S. Everett, "Improvement of the Excitation Source in the Narrow-Band Linear Prediction Vocoder", IEEE Trans. Acoust., Speech and Signal Proc., Vol. ASSP-33, pp.377-386, 1985). LPC inverse filtering can not produce a residual signal of absolutely flat magnitude spectrum mainly due to: a) the "cascade representation" of formats by the LPC filter 1/A(z), which results in the magnitudes of the resonant peaks to be dependent upon the pole locations of the 1/A(z) all-pole filter and b) the LPC quantisation noise. As a consequence, the LPC residual

signal is itself highly intelligible. Based on this observation the MG_j^n magnitudes are obtained by spectral sampling at the harmonic locations, ω_j^n, j=1,...., $\lfloor (P_n + 1)/2 \rfloor$, of a modified LPC synthesis filter, that is defined as follows:

$$MP(z) = \frac{G_N}{1 - G_R \sum_{i=1}^{p} \tilde{a}_i^n z^{-i}} \tag{32}$$

where, $\tilde{a}_i^n$, i=1,...,p represent the p quantised LPC coefficients of the nth coding frame and G_R and G_N are defined as follows:

$$G_R = \tilde{G}_R \sqrt{\prod_{i=1}^{p} (1 - K_i^n)^2} \tag{33}$$

and

$$G_N = \sqrt{\frac{\frac{1}{2P_n} \sum_{i=0}^{2P_n - 1} x_n^{rm}(i)^2}{\sum_{j=1}^{\lfloor (P_n+1)/2 \rfloor - 1} \left(\hat{MP}(\omega_j^n) H(\omega_j^n) \right)^2 / 2}} \tag{34}$$

where K_i^n , i=1,...,p are the reflection coefficients of the nth coding frame, $x_n^{rm}(i)$ represents a sequence of $2P_n$ speech samples centred in the middle of the nth coding frame from which the mean value is being calculated and removed, $\hat{MP}(\omega_j^n)$ and $H(\omega_j^n)$ represent the frequency response of the MP(z) and 1/A(z) filters respectively at the ω_j^n frequency. Notice that the $\hat{MP}(\omega_j^n)$ values are calculated assuming G_N=1. The $\tilde{G}_R$ parameter represents a constant whose value is set to 0.25.

Equation 32 defines a modified LPC synthesis filter with reduced feedback gain, whose frequency response consists of nearly equalised resonant peaks, the locations of which are very close to the LPC synthesis resonant locations. Furthermore, the value of the feedback gain G_R is controlled by the performance of the LPC model (i.e. it is proportional to the normalised LPC prediction error). In addition Equation 34 ensures that the energy of the reproduced speech signal is equal to the energy of the original speech waveform. Robustness is increased by computing the speech RMS value over two pitch periods.

Two alternative magnitude spectrum representation techniques are described below, which allow for better coding of the magnitude information and lead to a significant improvement in reconstructed speech quality.

The first of the alternative magnitude spectrum representations techniques is referred to below in the "Na amplitude system". The basic principle of this MG_j'' quantisation system is to represent accurately those MG_j'' values which correspond to the Na largest speech Short Term (ST) spectral envelope values. In particular, given the LPC coefficients of the nth coding frame, the ST magnitude spectrum envelope is calculated (i.e. sampled) at the harmonic frequencies ω_j'' and the locations lc(j), j=1,...,Na of the largest Na spectral samples are determined. These locations indicate effectively which of the $\lceil (P_n + 1)/2 \rceil - 1$ MG_j'' magnitudes are subjectively more important for accurate quantization. The system subsequently selects MG_{jn} j=lc(1),...,lc(Na) and Vector Quantizes these values. If the minimum pitch value is 17 samples. the number of non-DC MG_j'' amplitudes is equal to 8 and for this reason Na≤8. Two variations of the "Na-amplitudes system" were developed with equivalent performance and their block diagrams are depicted in Figure 33 (a) and (b) respectively.

i) Na-amplitudes system with Mean Normalization Factor. In this variation, a pitch segment of P_n residual samples $R^n(i)$, centered about the middle M_n of the nth coding frame is obtained and DFT transformed. The mean value of the spectral magnitudes MG_j'', j=1,..., $\lfloor (P_n + 1)/2 \rfloor$ is calculated as:

$$m = \frac{\sum_{j=1}^{\left\lceil \frac{P_n+1}{2} \right\rceil - 1} MG_j''}{\left\lceil \frac{P_n + 1}{2} \right\rceil - 1} \qquad (35)$$

m is quantized and then used as the normalization factor of the Na selected amplitudes MG''_j, j=lc(1),...,lc(Na). The resulting Na amplitudes are then vector quantized to MG''_j.

ii) Na-amplitudes system with RMS Normalization Factor. In this variation the RMS value of the pitch segment centered about the middle M_n of the nth coding frame, is calculated as:

$$g = \sqrt{\frac{\frac{1}{P_n}\sum_{i=0}^{P_n-1} R''(i)^2}{\frac{1}{2}\times\left(\left\lceil\frac{P_n+1}{2}\right\rceil-1\right)}} \qquad (36)$$

g is quantized and then used as the normalization factor of the Na selected amplitudes MG''_j, j=lc(1),...,lc(Na). These normalized amplitudes are then Vector Quantised to MG''_j. Notice that the P_n points DFT operation can be avoided in this case, since the magnitude spectrum of the pitch segment is calculated only at the Na selected harmonic frequencies ω''_j, j=lc(1),...,lc(Na).

In both cases the quantisation of the m and g factors, used to normalize the MG''_j values, is performed using an adaptive μ-law quantiser with a non-linear characteristic as:

$$c(A) = A_{max}\frac{\log_e(1+\mu|A|/A_{max})}{\log_e(1+\mu)}\mathrm{sgn}(A) \quad \text{with } \mu=255 \qquad (37)$$

This arrangement for the quantization of g or m extends the dynamic range of the coder to not less than 25dBs.

At the receiver end the decoder recovers the MG''_j magnitudes as $\hat{MG}''_j = MG'_j{}'' \times A$, j=lc(1),...,lc(Na). The remaining $\lceil(P_n+1)/2\rceil - Na - 1$ MG''_j values are set to a constant value A. (where A is either "m" or "g"). The block diagram of the adaptive μ-law quantiser is shown in Figure 34.

The second of the alternative magnitude spectrum representation techniques is referred to below as the "Variable Size Spectral Vector Quantisation (VS/SVQ)" system. Coding systems, which employ the general synthesis formula of Equation (1) to recover speech, encounter the problem of coding a variable length, pitch dependant spectral amplitude vector $\underline{MG}$. The "Na- amplitudes" MG_j'' quantisation schemes described in Figure 33 avoid this problem by Vector Quantising the minimum expected number of spectral amplitudes and by setting the rest of the MG_j'' amplitudes to a fixed value. However, such a partially spectrally flat excitation model has limitations in providing high recovered speech quality. Thus, in order to improve the output speech quality, the shape of the entire $\{ MG_j'' \}$ magnitude spectrum should be quantised. Various techniques have been proposed for coding $\{ MG_j'' \}$. Originally ADPCM has been used across the MG_j'' values associated to a specific coding frame. Also $\{ MG_j'' \}$ has been DCT transformed and coded differentially across successive MG_j'' magnitude spectra. However, these coding schemes are rather inefficient and operate with relatively high bit rates. The introduction of Vector Quantisation on the $\{ MG_j'' \}$ spectral amplitude vectors allowed for the development of Sinusoidal and Prototype Interpolation systems which operate at around 2.4 Kbits/sec. Two known $\{ MG_j'' \}$ VQ methods are described below which quantise a variable size (vs_n) input vector with a fixed size (fxs) codevector.

i) The first VQ method involves the transformation of the input vector to a fixed size vector followed by conventional Vector Quantisation. The inverse transformation on the quantised fixed size vector yields the recovered quantised $\underline{\hat{MG}''}$ vector. Transformation techniques which have been used include, Linear Interpolation, Band Limited Interpolation, All Pole modelling and Non-Square transformation. However, the overall distortion produced by this approach is the summation of the VQ noise and a component, which is introduced by the transformation process.

ii) The second VQ method achieves the direct quantisation of a variable input vector with a fixed size code vector. This is based in selecting only vs_n elements from each codebook vector, to form a distortion measure between a codebook vector and an input $\underline{MG^n}$ vector. Such a quantisation approach avoids the transformation distortion of the previous techniques mentioned in (i) and results in an overall distortion that is equal to the Vector Quantisation noise.

An improved VQ method will now be described which is referred to below as the Variable Size Spectral Vector Quantisation (VS/SVQ) scheme. This scheme was developed to take advantage of the underlying principle that the actual shape of the $\{ MG_j'' \}$ magnitude spectrum is defined by a minimum $\lceil (P_n + 1)/2 \rceil$ of equally spaced samples. If we consider the maximum expected pitch estimate P_{max}, then any $\{ MG_j'' \}$ spectral shape can be represented adequately by $\lceil (P_n + 1)/2 \rceil$ samples. This suggests that the fixed size fxs of the codebook vectors $\underline{S^i}$ representing the MG_j'' shapes should not be larger than $\lceil (P_n + 1)/2 \rceil$. Of course this also implies that given the $\lceil (P_n + 1)/2 \rceil$ samples of a codebook vector, the complete spectral shape, defined at any frequency, is obtained via an interpolation process.

Figure 35 highlights the VS/SVQ process. The codebook CBS having cbs fixed fxs dimension vectors S_j', j=1,...,fxs and i=1,...,cbs, where fxs is $\lceil (P_n + 1)/2 \rceil$, is used to quantise an input vector MG_j'', j=1,...,vs_n of dimension vs_n. Interpolation (in this case linear) is used on the $\underline{S^i}$ vectors to yield $\underline{S^{i\prime\prime}}$ vectors of dimension vs_n. The $\underline{S^i}$ to $\underline{S^{i\prime\prime}}$ interpolation process is given by:

$$S''(j) = S^i\left(\left\lfloor j\frac{fxs}{vs_n}\right\rfloor\right) + \left(j\frac{fxs}{vs_n} - \left\lfloor j\frac{fxs}{vs_n}\right\rfloor\right) \times \frac{S^i\left(\left\lceil j\frac{fxs}{vs_n}\right\rceil\right) - S^i\left(\left\lfloor j\frac{fxs}{vs_n}\right\rfloor\right)}{\left\lceil j\frac{fxs}{vs_n}\right\rceil - \left\lfloor j\frac{fxs}{vs_n}\right\rfloor} \qquad (38)$$

for i=1,...,cbs and j=1,...,vs_n

This process effectively defines $\underline{S''}$ spectral shapes at the ω_j'' frequencies of the MG_j'' vector. A distortion measure $D(\underline{S''}, \underline{MG''})$ is then defined between the $\underline{S''}$ and $\underline{MG''}$ vectors, and the codebook vector $\underline{S^l}$ that yields the minimum distortion is selected and its index I is transmitted. Of course in the receiver, Equation (38) is used to define $\underline{M\hat{G}''}$ from $\underline{S^l}$.

If we assume that $P_{max} \approx 120$ then fxs=60. However this value can be reduced to 50 without significant degradation by low pass filtering the signal synthesised from Equation (1). This is achieved by setting to zero all the harmonics MG_j'' in the region of 3.4 to 4.0KHz, in which case:

$$\left\lceil \frac{3400 \times p^n}{fs} \right\rceil = vs_n \qquad \text{if } vs_n \le 50 \qquad \textbf{(39)}$$

$$50 = vs_n \qquad \text{otherwise.}$$

and $vs_n \le fxs$.

Amplitude vectors, obtained from adjacent residual frames, exhibit significant redundancy, which can be removed by means of backward prediction. Prediction is performed on a harmonic basis i.e. the amplitude value of each harmonic MG_j'' is predicted from the amplitude value of the same harmonic in previous frames i.e. $MG_j''^{-1}$. A fixed linear predictor $\underline{M\tilde{G}''} = b \times \underline{M\hat{G}''^{-1}}$ may be incorporated in the VS/SVQ system, and the resulting DPCM structure is shown in Figure 36 (differential VS/SVQ, (DVS/SVQ)). In particular, error vectors are formed as the difference between the original spectral amplitudes MG_j'' and their predicted ones $M\tilde{G}_j^n$, i.e.:

$$E_j'' = MG_j'' - M\tilde{G}_j'' \text{ for } j=1,\ldots,vs_n.$$

where the predicted spectral amplitudes $M\tilde{G}_j''$ are given as:

$$M\tilde{G}_j'' = \begin{cases} b \times M\hat{G}_j''^{-1} & when \quad V_{n-1} = 1 \\ 0 & when \quad V_n = 0 \end{cases} \qquad \text{for } 1 \le j \le vs_{n-1} \qquad \textbf{(40)}$$

and

$$\widetilde{MG}_j^n = \frac{1}{vs_{n-1}} \sum_{k=1}^{vs_{n-1}} \widetilde{MG}_k^n . \qquad \text{for } vs_{n-1} < j \le vs_n \qquad (41)$$

Furthermore the quantised spectral amplitudes $\hat{MG}_j^n$ are given as:

$$\hat{MG}_j^n = \begin{cases} \widetilde{MG}_j^n + \hat{E}_j^n & for \quad 1 \le j \le vs_n \\ \dfrac{1}{vs_n} \sum_{k=1}^{vs_n} \hat{MG}_k^n & for \quad vs_n < j < \left\lceil \dfrac{P_n + 1}{2} \right\rceil \end{cases} \qquad (42)$$

where $\hat{E}_j^n$ denotes the quantised error vector.

The quantisation of the E_j^n $1 \le j \le vs_n$ error vector incorporates Mean Removal and Gain Shape Quantisation techniques, using the hierarchical VQ structure of Figure 36.

A weighted Mean Square Error is used in the VS/SVQ stage of the system. The weighting function is defined as the frequency response of the filter: $W(z) = 1 / A_n(z/\gamma)$, where $A_n(z)$ is the short-term linear prediction filter and γ is a constant, defined as γ=0.93. Such a weighting function that is proportional to the short-term envelope spectrum, results in substantially improved decoded speech quality. The weighting function W_j^n is normalised so that:

$$\sum_{j=1}^{vs_n} W_j^n = 1 \qquad (43)$$

The pdf of the mean value of $\underline{E^n}$ is very broad and, as a result, the mean value differs widely from one vector to another. This mean value can be regarded as statistically independent of the variation of the shape of the error vector $\underline{E^n}$ and thus, can be quantised separately without paying a substantial penalty in compression efficiency. The mean value of an error vector is calculated as follows:

$$M = \sum_{j=1}^{vs_n} W_j^n \times E_j^n \qquad (44)$$

M is Optimum Scalar Quantised to $\hat{M}$ and is then removed from the original error vector to form $\underline{Erm^n} = (\underline{E^n} - \hat{M})$. The overall quantization distortion is attributed to the quantization

of the "Mean Removed" error vectors ($\underline{Erm}^n$), which is performed by a Gain-Shape Vector Quantiser.

The objective of the Gain-Shape VQ process is to determine the gain value $\hat{C}$ and the shape vector $\hat{\underline{S}}$ so as to minimise the distortion measure:

$$D(\underline{Erm}^n, \hat{G} \times \hat{\underline{S}}) = \sum_{j=1}^{vs_n} W_j^n \left[Erm_j^n - \hat{G} \times \hat{S}_j \right]^2 \qquad (45)$$

A gain optimised VQ search method, similar to techniques used in CELP systems, is employed to find the optimum $\hat{G}$ *and* $\hat{\underline{S}}$. The shape Codebook (CBS) of vectors $\underline{S}^i$ is searched first to yield an index I, which maximises the quantity:

$$Q(i) = \frac{\left(\sum_{j=1}^{vs_n} W_j^n Erm_j^n(i) S_j^{i} \right)^2}{\sum_{j=1}^{vs_n} W_j^n S_j^{i\,2}} \qquad \text{for i=1,...,cbs} \qquad (46)$$

where cbs is the number of codevectors in the CBS. The optimum gain value is defined as:

$$G = \frac{\sum_{j=1}^{vs_n} W_j^n Erm_j^n S_j^{I}}{\sum_{j=1}^{vs_n} W_j^n S_j^{I\,2}} \qquad (47)$$

and is Optimum Scalar Quantised to $\hat{G}$.

During shape quantisation the principles of VS/SVQ are employed, in the sense that the $\underline{S}^{i}$, vs_n size vectors are produced using Linear Interpolation on fxs size codevectors $\underline{S}^i$. Both trained and randomly generated shape CBS codebooks were investigated. Although $\underline{Erm}^n$ has noise-like characteristics, systems using randomly generated shape codebooks resulted in unsatisfactory muffled decoded speech and were inferior to systems employing trained shape codebooks.

A closed-loop joint predictor and VQ design process was employed to design the CBS codebook, the optimum scalar quantisers CBM and CBG of the mean M and gain G values respectively, and also to define the prediction coefficient b of Figure 36. In particular, the following steps take place in the design process.

STEP A0 (k=0). Given a training sequence of MG_j^n the predictor b^0 is calculated in an open loop fashion (i.e. $\tilde{MG}_j^n = b \times MG_j^{n-1}$ for $1 \le j < \lceil (P_n + 1)/2 \rceil$ when $V_{n-1}=1$, or $\tilde{MG}_j^n = 0$ elsewhere). Furthermore, the CBM^0 mean, CBG^0 gain and CBS^0 shape codebooks are designed independently and again in an open loop fashion using unquantized $\underline{E}^n$. In particular:

a) Given a training sequence of error vectors $\underline{E}^{n\,0}$, the mean value of each $\underline{E}^{n\,0}$ is calculated and used in the training process of an Optimum Scalar Quantiser (CBM^0).

b) Given a training sequence of error vectors $\underline{E}^{n\,0}$ and the CBM^0 mean quantiser, the mean value of each error vector is calculated, quantised using the CBM^0 quantiser and removed from the original error vectors $\underline{E}^{n\,0}$ to yield a sequence of "Mean Removed" training vectors $\underline{Erm}^{n\,0}$.

c) Given a training sequence of $\underline{Erm}^{n\,0}$ vectors, each "Mean Removed" training vector is normalised to unit power (i.e. is divided by the factor $G = \sqrt{\sum_{i=1}^{m_n} W_j^n \left(Erm_j^{n\,0}\right)^2}$), linear interpolated to fxs points, and then used in the training process of a conventional Vector Quantiser of fxs dimension. (CBS^0).

d) Given a training sequence of $\underline{Erm}^{n\,0}$ vectors and the CBS^0 shape codebook, each "Mean Removed" training vector is encoded using Equations 46 and 47 and the value G of Equation 47 is used in the training process of an Optimum Scalar Quantiser (CBG^0).

k is set to 1 (k=1).

STEP A1 Given a training sequence of MG_j and the mean, gain and shape codebooks of the previous k-1 iterations (i.e. CBM^{k-1}, CBG^{k-1}, CBS^{k-1}), the optimum prediction coefficient b^k is calculated.

STEP A2 Given a training sequence of MG_j , an optimum prediction coefficient b^k and CBM^{k-1}, CBG^{k-1}, CBS^{k-1}, a training sequence of error vectors $\underline{E''}^k$ is formed, which is then used for the design of new mean, gain and shape codebooks (i.e. CBM^k, CBG^k, CBS^k).

STEP A3 The performance of the kth iteration quantization system (i.e. b^k, CBM^k, CBG^k, CBS^k) is evaluated and compared against the quantization system of the previous iteration (i.e. b^{k-1} , CBM^{k-1}, CBG^{k-1}, CBS^{k-1}). If the quantization distortion converges to a minimum, the quantization design process stops. Otherwise, k=k+1 and steps A1, A2 and A3 are repeated.

The performance of each quantizer (i.e. b^k , CBM^k, CBG^k, CBS^k) has been evaluated using subjective tests and a LogSegSNR distortion measure, which was found to reflect the subjective performance of the system.

The design for the Mean-Shape-Gain Quantiser used in **STEP A2** is performed using the following two steps :

STEP B1 Given a training sequence of error vectors $\underline{E''}^k$, the mean value of each $\underline{E''}^k$ is calculated and used in the training process of an Optimum Scalar Quantiser (CBM^k).

STEP B2 Given a training sequence of error vectors $\underline{E''}^k$ and the CBM^k mean quantizer, the mean value of each residual vector is calculated, quantized and removed from the original residual vectors $\underline{E''}^k$ to yield a sequence of "Mean Removed" training vectors $\underline{Erm''}^k$, which are then used as the training data in the design of an optimum Gain Shape Quantizer (CBG^k and CBS^k). This involves steps C1 - C4 below. (The quantization design process is performed under the assumption of any independent

gain shape quantiser structure, i.e. an input error vector $\underline{Emr}^n$ can be represented by any possible combination of $\underline{S}^i$ codebook shape vectors and $\hat{G}$ gain quantizer levels.)

STEP C1 (v=0). Given a training sequence of vectors $\underline{Erm}^{n\ k}$ and an initial $CBG^{k,0}$ and $CBS^{k,0}$ gain and shape codebooks respectively, compute the overall average distortion distance $D_{k,0}$ as in Equation 44. Set v equal to 1 (v=1).

STEP C2 Given a training sequence of vectors $\underline{Erm}^{n\ k}$ and the $CBG^{k,v-1}$ gain codebook from the previous iteration, compute the new shape codebook $CBS^{k,v}$ which minimises the VQ distortion measure. Notice that the optimum $CBS^{k,v}$ shape codebook is obtained when the distortion measure of Equation (44) is a minimum and this is achieved in $M1_{k,v}$ iterations.

STEP C3 Given a training sequence of vectors $\underline{Erm}^{n\ k}$ and the $CBS^{k,v}$ shape codebook, compute a new gain quantiser $CBG^{k,v}$, which minimise the distortion measure of Equation (44). This optimum $CBG^{k,v}$ gain quantiser is obtained when the distortion measure of Equation (44) is a minimum and this is achieved in $M2_{k,v}$ iterations.

STEP C4 Given a training sequence of vectors $\underline{Erm}^{n\ k}$ and the shape and gain codebooks $CBS^{k,v}$ and $CBG^{k,v}$, compute the average overall distortion measure. If $(D_{k,v-1}-D_{k,v})/D_{k,v}<\varepsilon$ stop. Otherwise, v=v+1 and go back to **STEP C2**.

The centroids $S_{i,u}^{k,v,m}$, i=1,...,cbs and u=1,...,fxs of the shape Codebook $CBS^{k,v,m}$, are updated during the mth iteration performed in **STEP C2** (m=1,....$M1_{k,v}$) as follows:

$$S_{i,u}^{k,v,m} = C_{u,j,n} \frac{\sum_{n:\underline{Erm}^k \in Q_i} \left(NC_{u,j,n} + \tilde{C}_{u,j,n} NC_{u,j',n}\right)}{\sum_{n:\underline{Erm}^k \in Q_i} \left(DC_{u,j,n} + \tilde{C}_{u,j,n} DC_{u,j',n}\right)} \qquad (48)$$

where $DC_{j,n,u} = W_j^n \left(G_{j_n}^{k,v-1} \times f_{u,j,n}\right)^2$.

$$NC_{u,j,n} = W_j^n G_{j_n}^{k,v-1} f_{u,j,n} \left(Erm_j^{n\,k} - G_{j_n}^{k,v-1} S_{j,u^*(u,j,n)}^{k,v,m-1} (1 - f_{u,j,n})\right).$$

$$f_{u,j,n} = 1 - \left| j\frac{fxs}{vs_n} - u \right|,$$

$$C_{u,j,n} = \begin{cases} 1 & if \quad f_{u,j,n} \le 1 \\ 0 & if \quad f_{u,j,n} > 1 \end{cases},$$

$$\tilde{C}_{u,j,n} = \begin{cases} 1 & if \quad \left| j\frac{fxs}{vs_n} - u \right| + 1 \le \frac{fxs}{vs_n} \\ 0 & if \quad \left| j\frac{fxs}{vs_n} - u \right| + 1 > \frac{fxs}{vs_n} \end{cases};$$

$$u''(u,j,n) = \begin{cases} u+1 & if \quad u < j\frac{fxs}{vs_n} \\ u-1 & if \quad u \ge j\frac{fxs}{vs_n} \end{cases} \quad \text{and}$$

$$j' = \begin{cases} j-1 & if \quad u < j\frac{fxs}{vs_n} \\ j+1 & if \quad u \ge j\frac{fxs}{vs_n} \end{cases}$$

Q_i denotes the cluster of $\underline{Erm^n}^k$ error vectors which are quantised to the $\underline{S_i^{k,v,m-1}}$ codebook shape vector, cbs represents the total number of shape quantisation levels, J_n represents the $CBG^{k,v-1}$ gain codebook index which encodes the $\underline{Erm^n}^k$ error vector and $1 \le j \le vs_n$.

The gain centroids, $G_i^{k,v,m}$, i=1,...,cbg of the $CBG^{k,v,m}$ gain quantiser, which are computed during the mth iteration in **STEP C3** (m=1.....$M2_{k,v}$), are given as:

$$G_i^{k,v,m} = \frac{\sum_{n:\underline{Erm}^{n\,k} \in D_i} \left(\sum_{j=1}^{vs_n} Erm_j^{n\,k} S_{I_n}^{j\,k,v} W_j^n \right)}{\sum_{n:\underline{Erm}^{n\,k} \in D_i} \left(\sum_{j=1}^{vs_n} \left(S_{I_n}^{j\,k,v} \right)^2 W_j^n \right)} \qquad (49)$$

where D_i denotes the cluster of $\underline{Erm}^{n\,k}$ error vectors which are quantised to the $G_i^{k,v,m-1}$ gain quantiser level, cbg represents the total number of gain quantisation levels, I_n represents the $CBS^{k,v}$ shape codebook index which encodes the $\underline{Erm}^{n\,k}$ error vector and $1 \le j \le vs_n$.

The above employed design process is applied to obtain the optimum shape codebook CBS, optimum gain and mean quantizers, CBG and CBM and the optimum prediction coefficient b which was finally set to b=0.35.

Process VII calculates the energy of the residual signal. The LPC analysis performed in Process II provides the prediction coefficients a_i $1 \le i \le p$ and the reflection coefficients k_i $1 \le i \le p$. On the other hand, the Voiced/Unvoiced classification performed in Process I provides the short term autocorrelation coefficient for zero delay of the speech signal (R0) for the frame under consideration. Hence, the Energy of the residual signal E_n value is given as:

$$E_n = \frac{1}{M} RO \prod_{i=1}^{p} (1 - K_i)^2 \qquad (50)$$

The above expression represents the minimum prediction error as it is obtained from the Linear Prediction process. However, because of quantization distortion the parameters of the LPC filter used in the coding-decoding process are slightly different from the ones that achieve minimum prediction error. Thus, Equation (50) gives a good approximation of the residual signal energy with low computational requirements. The accurate E_n value can be given as:

$$E_n = \frac{1}{M}\sum_{i=0}^{M-1} R''(i)^2 \qquad (51)$$

The resulting $\sqrt{E_n}$ is then Scalar Quantised using an adaptive μ-law quantised arrangement similar to the one depicted in Figure 34. In the case where more than one $\sqrt{E_n}$ are used in the system i.e. the energy E_n is calculated for a number of subframes then $E_{n,\xi}$ is given by the general equation:

$$E_{n,\xi} = \frac{1}{M_s}\sum_{i=0}^{M_s-1} R''(i+\xi M_s)^2 \qquad 0 \le \xi \le \Xi \qquad (52)$$

Notice that when $\Xi = 1, M_s = M$ and for $\Xi = 4, M_s = M/4$.

CLAIMS

1. A speech synthesis system in which a speech signal is divided into a series of frames, and each frame is converted into a coded signal including a voiced/unvoiced classification and a pitch estimate, wherein a low pass filtered speech segment centred about a reference sample is defined in each frame, a correlation value is calculated for each of a series of candidate pitch estimates as the maximum of multiple crosscorrelation values obtained from variable length speech segments centred about the reference sample, the correlation values are used to form a correlation function defining peaks, and the locations of the peaks are determined and used to define a pitch estimate.

2. A system according to claim 1, wherein the pitch estimate is defined using an iterative process.

3. A system according to claim 1 or 2, wherein a single reference sample may be used, centred with respect to the respective frame.

4. A system according to claim 1 or 2, wherein multiple pitch estimates are derived for each frame using different reference samples, the multiple pitch estimates being combined to define a combined pitch estimate for the frame.

5. A system according to any preceding claim, wherein the pitch estimate is modified by reference to a voiced/unvoiced status and/or pitch estimates of adjacent frames to define a final pitch estimate.

6. A system according to any preceding claim, wherein the correlation function is clipped using a threshold value, remaining peaks being rejected if they are adjacent to larger peaks.

7. A system according to claim 6, wherein peaks are selected which are larger that either adjacent peak and peaks are rejected if they are smaller than a following peak by more than a predetermined factor.

8. A system according to any preceding claim, wherein the pitch estimation procedure is based on a least squares error algorithm.

9. A system according to claim 8, wherein the pitch estimation algorithm defines the pitch valve as a number whose multiples best fit the correlation function peak locations.

10. A system according to any preceding claim, wherein possible pitch values are limited to integral numbers which are not consecutive, the increment

between two successive numbers being proportional to a constant multiplied by the lower of those two numbers.

11. A speech synthesis system in which a speech signal is divided into a series of frames, and each frame is converted into a coded signal including pitch segment magnitude spectral information, a voiced/unvoiced classification, and a mixed voiced classification which classifies harmonics in the magnitude spectrum of voiced frames as strongly voiced or weakly voiced, wherein a series of samples centred on the middle of the frame are windowed to form a data array which is Fourier transformed to produce a magnitude spectrum, a threshold value is calculated and used to clip the magnitude spectrum, the clipped data is searched to define peaks, the locations of peaks are determined, constraints are applied to define dominant peaks, and harmonics not associated with a dominant peak are classified as weakly voiced.

12. A system according to claim 11, wherein peaks are located using a second order polynomial

13. A system according to claim 11 or 12, wherein the samples are Hamming windowed.

14. A system according to claim 11, 12 or 13, wherein the threshold value is calculated by identifying the maximum and minimum magnitude spectrum values and defining the threshold as a constant multiplied by the difference between the maximum and minimum values.

15. A system according to any one of claims 11 to 14, wherein peaks are defined as those values which are greater than the two adjacent values, a peak being rejected from consideration if neighbouring peaks are of a similar magnitude or if there are spectral magnitudes in the same range of greater magnitude.

16. A system according to any one of claims 11 to 15, wherein a harmonic is considered as not being associated with a dominant peak if the difference between two adjacent peaks is greater than a predetermined threshold value.

17. A system according to any one of claims 11 to 16, wherein the spectrum is divided into bands of fixed width and a strongly/weakly voiced classification is assigned for each band.

18. A system according to any one of claims 11 to 17, wherein the frequency range is divided into two or more bands of variable width, adjacent bands being

separated at a frequency selected by reference to the strongly/weakly voiced classification of harmonics.

19. A system according to claim 17 or 18, wherein the lowest frequency band is regarded as strongly voiced, whereas the highest frequency band is regarded as weakly voiced.

20. A system according to claim 19, wherein the event that a current frame is voiced, and the following frame is unvoiced, further bands within the current frame will be automatically classified as weakly voiced.

21. A system according to claim 19 or 20, wherein the strongly/weakly voiced classification is determined using a majority decision rule on the strongly/weakly voiced classification of those harmonics which fall within the band in question.

22. A system according to claim 21, wherein, if there is no majority, alternate bands are alternately assigned strongly voiced and weakly voiced classifications.

23. A speech synthesis system in which a speech signal is divided into a series of frames, each frame is defined as voiced or unvoiced, each frame is converted into a coded signal including a pitch period value, a frame voiced/unvoiced classification and, for each voiced frame, a mixed voiced spectral band

classification which classifies harmonics within spectral bands as either strongly or weakly voiced, and the speech signal is reconstructed by generating an excitation signal in respect of each frame and applying the excitation signal to a filter, wherein for each weakly voiced spectral band, an excitation signal is generated which includes a random component in the form of a function which is dependent upon the respective pitch period value.

24. A system according to claim 23, wherein the spectrum is divided into bands and a strongly/weakly voiced classification is assigned to each band.

25. A system according to claim 23 or 24, wherein the random component is introduced by reducing the amplitude of harmonic oscillators assigned the weakly voiced classification, disturbing the oscillator frequencies such that the frequency is no longer a multiple of the fundamental frequency, and then adding further random signals.

26. A system according to claim 25, wherein the phase of the oscillators is randomised.

27. A speech synthesis system in which a speech signal is divided into a series of frames, and each voiced frame is converted into a coded signal including a pitch period value LPC coefficients and pitch segment spectral magnitude

information, wherein the spectral magnitude information is quantized by sampling the LPC short term magnitude spectrum at harmonic frequencies, the locations of the largest spectral samples are determined to identify which of the magnitudes are relatively more important for accurate quantization, and the magnitudes so identified are selected and vector quantized.

28. A system according to claim 27, wherein a pitch segment of P_n LPC residual samples is obtained, where P_n is the pitch period value of the nth frame, the pitch segment is DFT transformed, the mean value of the resultant spectral magnitudes is calculated, the mean value is quantized and used as a normalisation factor for the selected magnitudes, and the resulting normalised amplitudes are quantized.

29. A system according to claim 27, wherein the RMS value of the pitch segment is calculated, the RMS value is quantized and used as a normalisation factor for the selected magnitudes, and the resulting normalised amplitudes are quantized.

30. A system according to any one of claims 27 to 29, wherein , at the receiver, the selected magnitudes are recovered, and each of the other magnitude values is reproduced as a constant value.

31. A speech synthesis system in which a variable size input vector of coefficients to be transmitted to a receiver for the reconstruction of a speech signal is vector quantized using a codebook defined by vectors of fixed size, the codebook vectors of fixed size are obtained from variable sized training vectors and an interpolation technique which is an integral part of the codebook generation process, codebook vectors are compared to the variable sized input vector using the interpolation process, and an index associated with the codebook entry with the smallest difference from the comparison is transmitted, the index being used to address a further codebook at the receiver and thereby derive an associated fixed size codebook vector, and the interpolation process being used to recover from the derived fixed sized codebook vector an approximation of the variable sized input vector.

32. A system according to claim 31, wherein the interpolation process is linear, and for an input vector of given dimension, the interpolation process is applied to produce from the codebook vectors a set of vectors of that given dimension, a distortion measure is then derived to compare the interpolated set of vectors and the input vector, and the codebook vector is selected which yields the minimum distortion.

33. A system according to claim 32, wherein the dimension of the vectors is reduced by taking into account only the harmonic amplitudes within an input bandwidth range.

34. A system according to claim 33, wherein the remaining amplitudes are set to a constant value.

35. A system according to claim 34, wherein the constant value is equal to the mean value of the quantized amplitudes.

36. A system according to any one of claims 31 to 35, wherein redundancy between amplitude vectors obtained from adjacent residual frames is removed by means of backward prediction.

37. A system according to claim 36, wherein the backward prediction is performed on a harmonic basis such that the amplitude value of each harmonic of one frame is predicted from the amplitude value of the same harmonic in the previous frame or frames.

38. A speech synthesis system in which a speech signal is divided into a series of frames, each frame is converted into a coded signal including an estimated pitch period, an estimate of the energy of a speech segment the duration of which

is a function of the estimated pitch period, and LPC filter coefficients defining an LPC spectral envelope, and a speech signal of related power to the power of the input speech signal is reconstructed by generating an excitation signal using spectral amplitudes which are defined from a modified LPC spectral envelope sampled at harmonic frequencies defined by the pitch period.

39. A system according to claim 38, wherein the magnitude values are obtained by spectrally sampling a modified LPC synthesis filter characteristic at the harmonic locations related to the pitch period.

40. A system according to claim 39, wherein the modified LPC synthesis filter has reduced feed back gain and a frequency response which consists of equalised resonant peaks, the locations of which are close to the LPC synthesis resonant locations.

41. A system according to claim 40, wherein the value of the feed back gain is controlled by the performance of the LPC model such that it is related to the normalised LPC prediction error.

42. A system according to any one of claims 38 to 41, wherein the energy of the reproduced speech signal is equal to the energy of the original speech waveform.

43. A speech synthesis system in which a speech signal is divided into a series of frames, each frame is converted into a coded signal including LPC filter coefficients and at least one parameter associated with a pitch segment magnitude, and the speech signal is reconstructed by generating two excitation signals in respect of each frame, each pair of excitation signals comprising a first excitation signal generated on the basis of the pitch segment magnitude parameter or parameters of one frame and a second excitation signal generated on the basis of the pitch segment magnitude parameter or parameters of a second frame which follows and is adjacent to the said one frame, applying the first excitation signal to a first LPC filter the characteristics of which are determined by the LPC filter coefficients of the said one frame and applying the second excitation signal to a second LPC filter the characteristics of which are determined by the LPC filter coefficients of the said second frame, and weighting and combining the outputs of the first and second LPC filters to produce one frame of a synthesised speech signal.

44. A system according to claim 43, wherein the first and second excitation signals include the same phase function and different phase contributions from the two LPC filters.

45. A system according to claim 44, wherein the outputs of the first and second LPC filters are weighted by half a window function such that the magnitude of the output of the first filter is decreasing with time and the magnitude of the output of the second filter is increasing with time.

46. A speech coding system which operates on a frame by frame basis, and in which information is transmitted which represents each frame as either voiced or unvoiced and, for each voiced frame, represents that frame by a pitch period value, quantized magnitude spectral information, and LPC filter coefficients, the received pitch period value and magnitude spectral information being used to generate residual signals at the receiver which are applied to LPC speech synthesis filters the characteristics of which are determined by the transmitted filter coefficients, wherein each residual signal is synthesised according to sinusoidal mixed excitation synthesis process, and a recovered speech signal is derived from the residual signals.

47. A speech synthesis system substantially as hereinbefore described with reference to the accompany drawings.

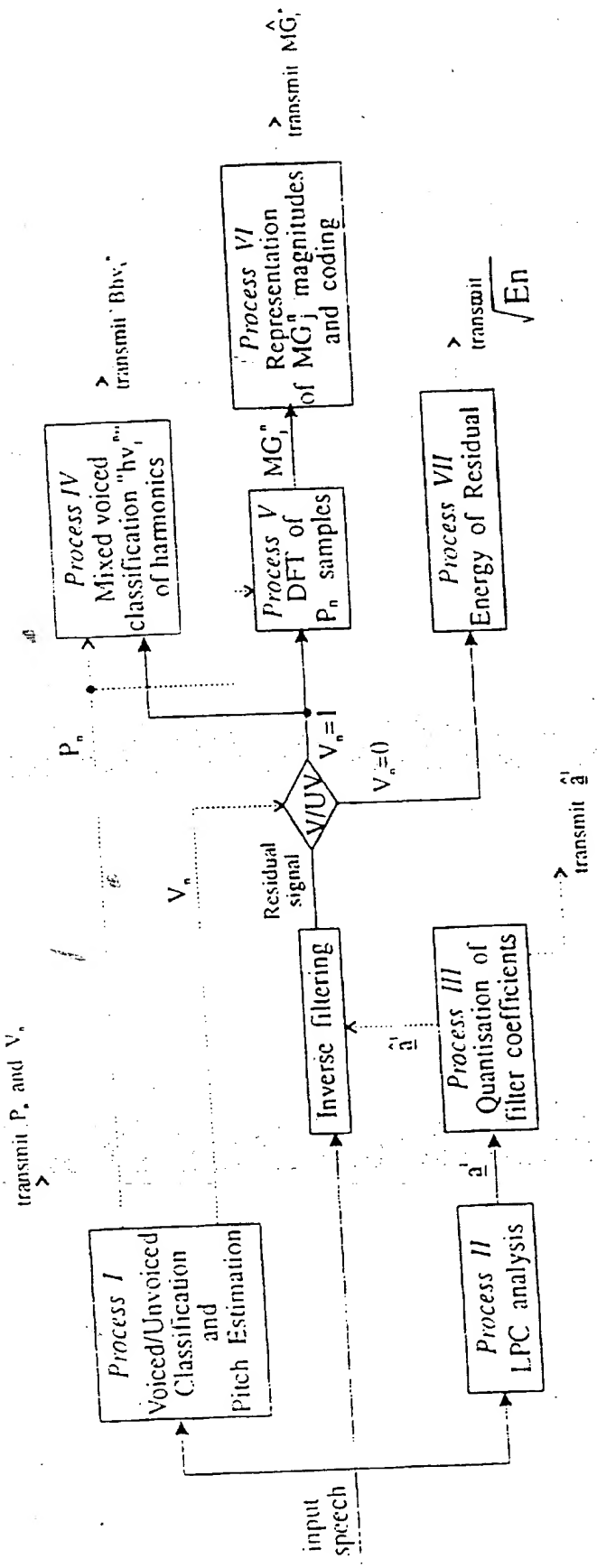

input speech
Process I Voiced/Unvoiced Classification and Pitch Estimation
Process II LPC analysis
Process III Quantisation of filter coefficients
Inverse filtering
Residual signal
V/UV
V_n=1
V_n=0
Process IV Mixed voiced classification "hv_i" of harmonics
Process V DFT of P_n samples
Process VI Representation of MG_j^n magnitudes and coding
Process VII Energy of Residual
transmit P_n and V_n
transmit $\hat{a}^1$
transmit $\hat{hv}_i$
transmit $\hat{MG}_j^n$
transmit $\sqrt{En}$

FIG.1

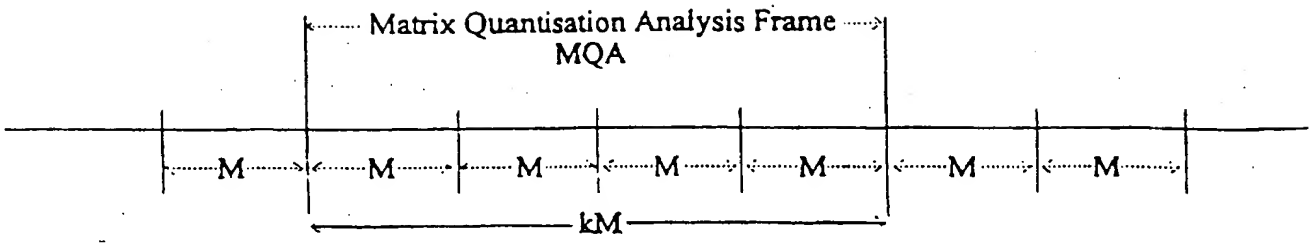


FIG. 2

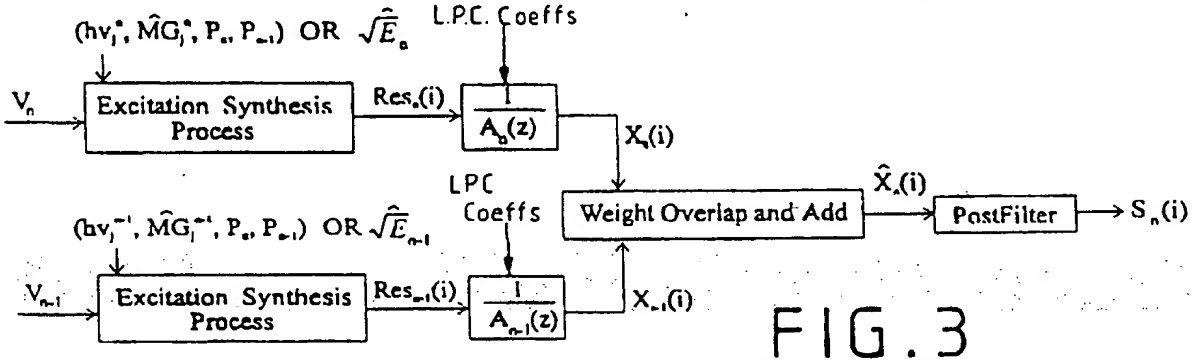


FIG. 3

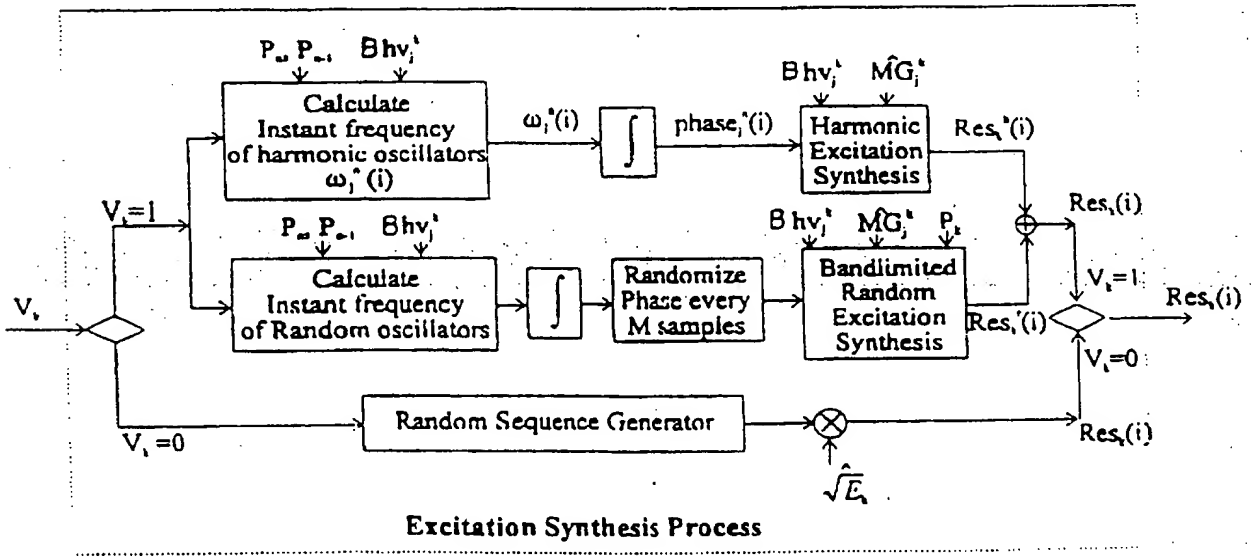


FIG. 4

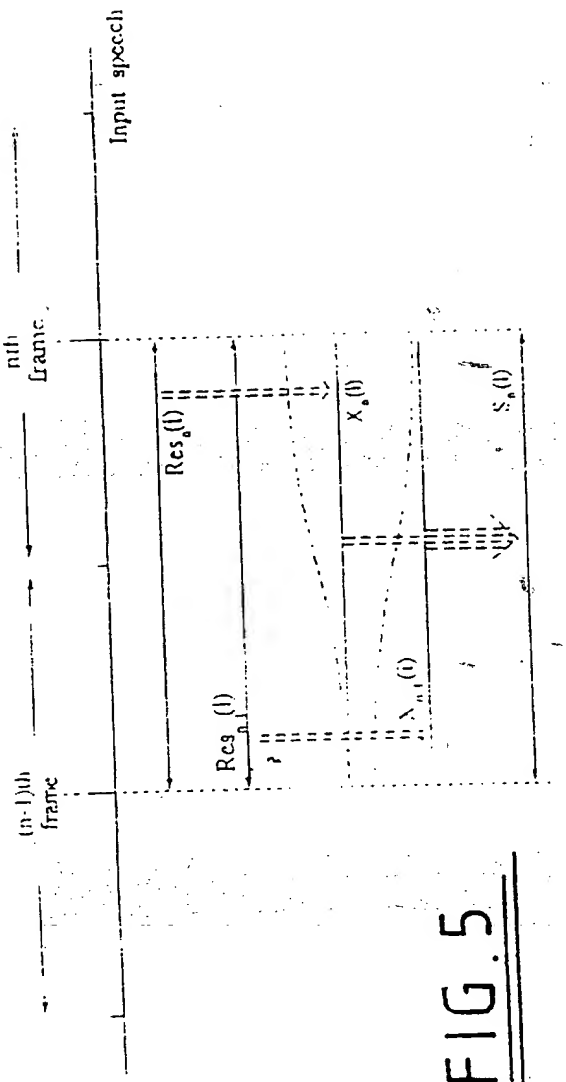

Input speech
nth frame
(n-1)th frame
$Res_n(l)$
$Res_{n-1}(l)$
$X_n(l)$
$X_{n-1}(l)$
$S_n(l)$

FIG. 5

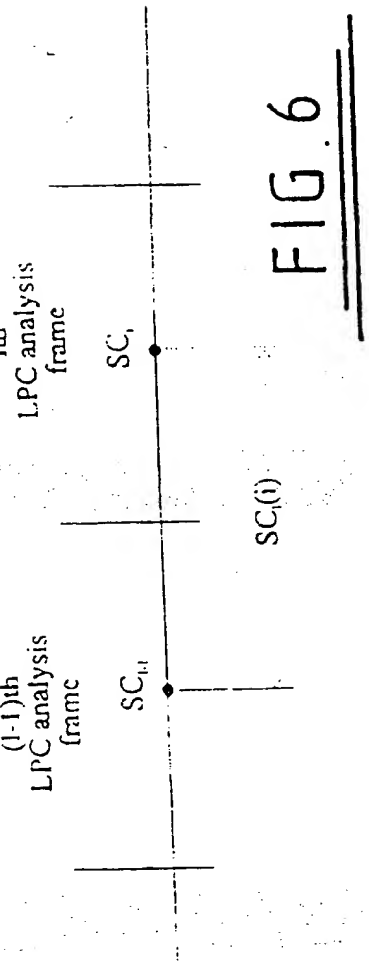

lth LPC analysis frame
(l-1)th LPC analysis frame
SC_l
SC_{l-1}
$SC_l(i)$

FIG. 6

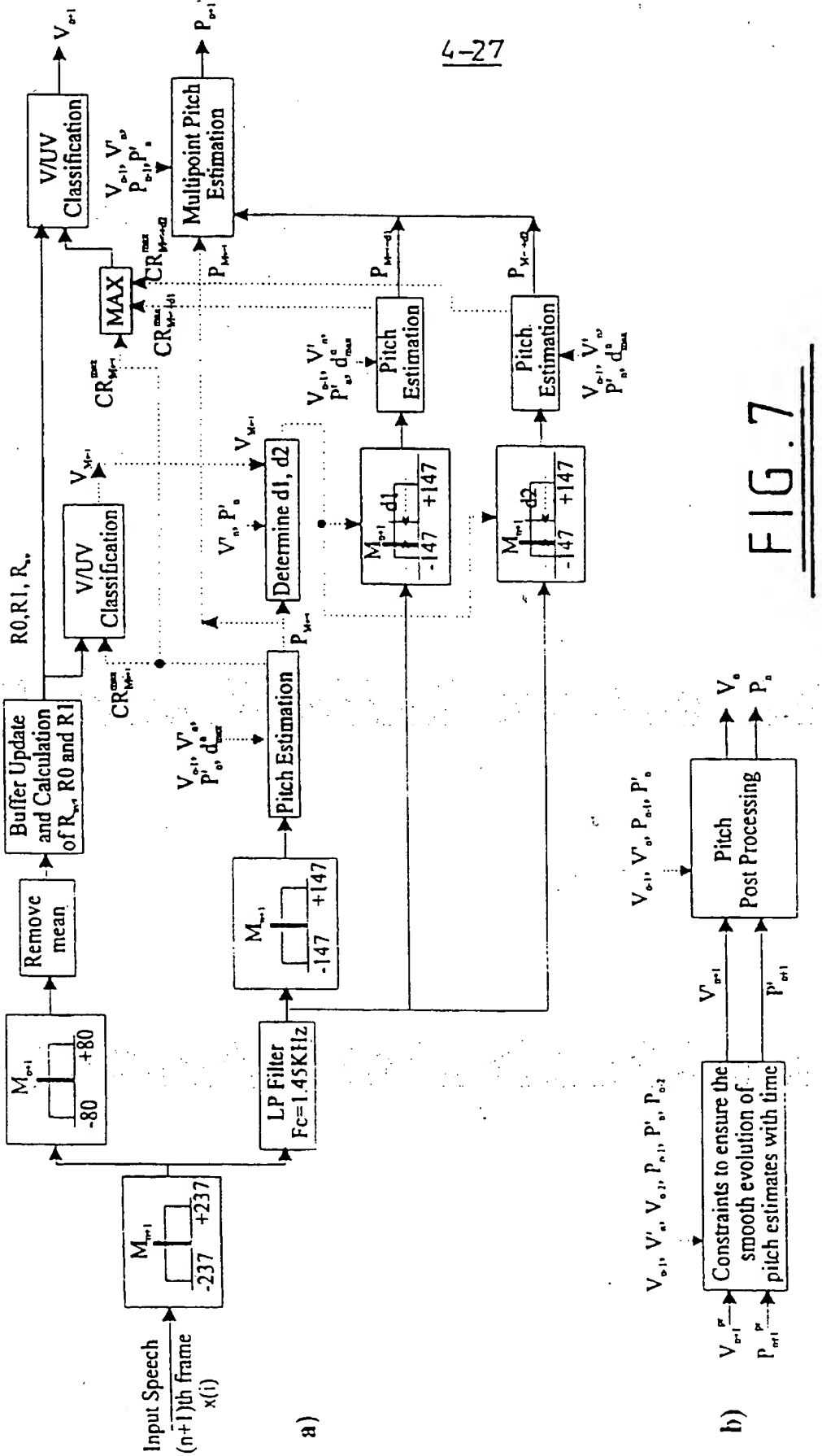

Input Speech
(n+1)th frame
x(i)
-237 +237
-80 +80
Remove mean
Buffer Update and Calculation of R, R0 and R1
R0,R1, R
V/UV Classification
V/UV Classification
MAX
Multipoint Pitch Estimation
Pitch Estimation
Determine d1, d2
Pitch Estimation
Pitch. Estimation
-147 +147
-147 +147
LP Filter
Fc=1.45KHz
-147 +147
a)
Constraints to ensure the smooth evolution of pitch estimates with time
Pitch Post Processing
b)

FIG. 7

5-27

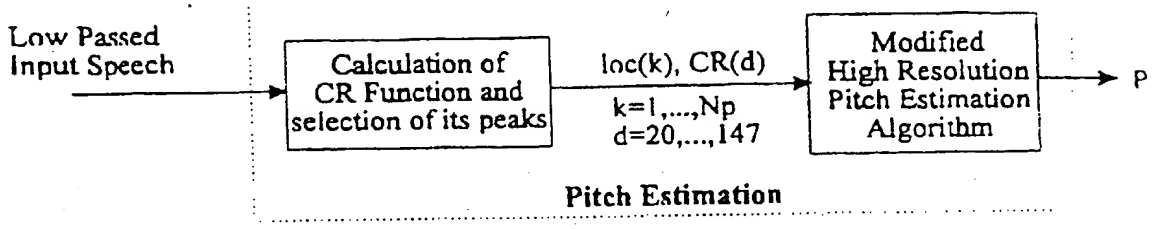


FIG. 8

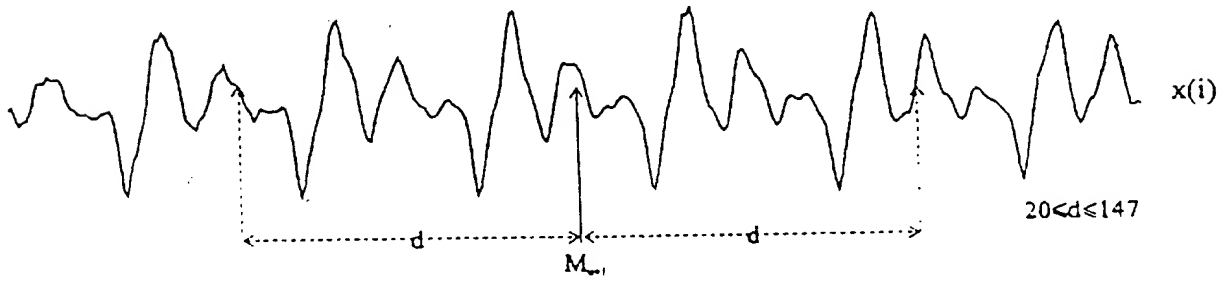


FIG. 9

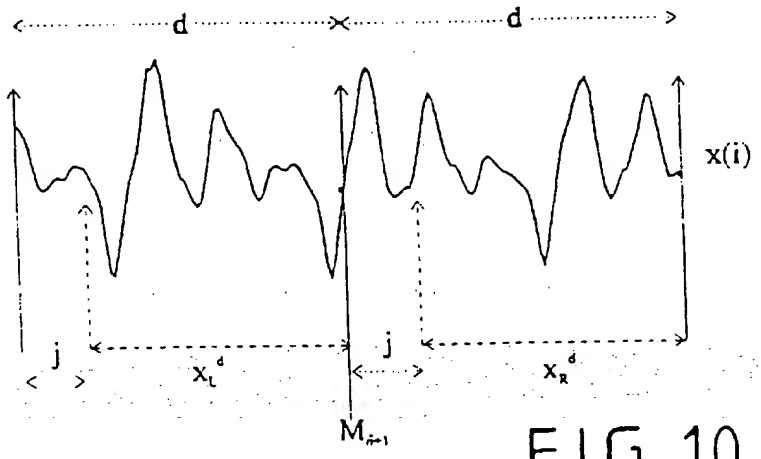
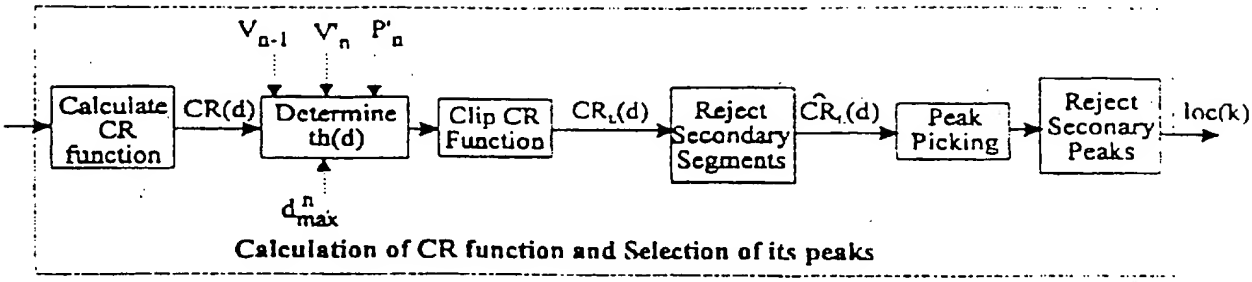


FIG. 10

d	20	30	35	37	40	90	147
f(d)	0	1	3	5	$d-\left\lfloor\frac{d\times 6}{7}\right\rfloor-1$	14	

FIG. 11



Calculation of CR function and Selection of its peaks

FIG. 12

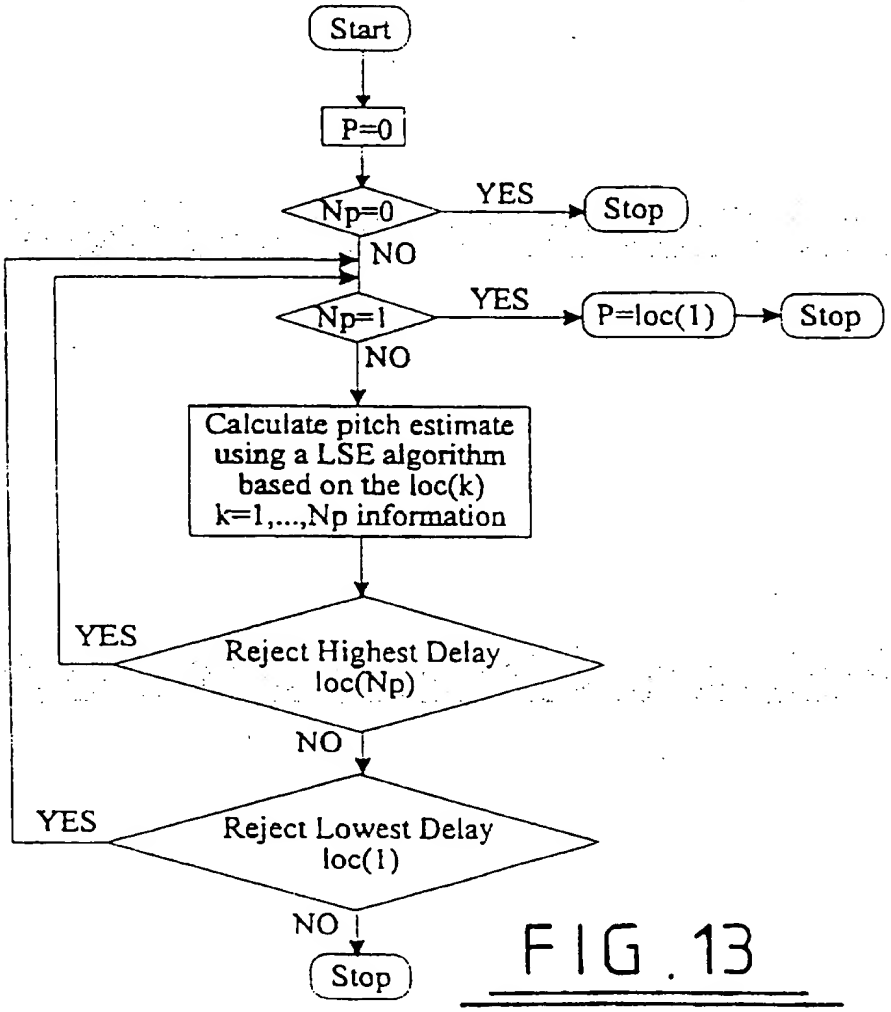


FIG. 13

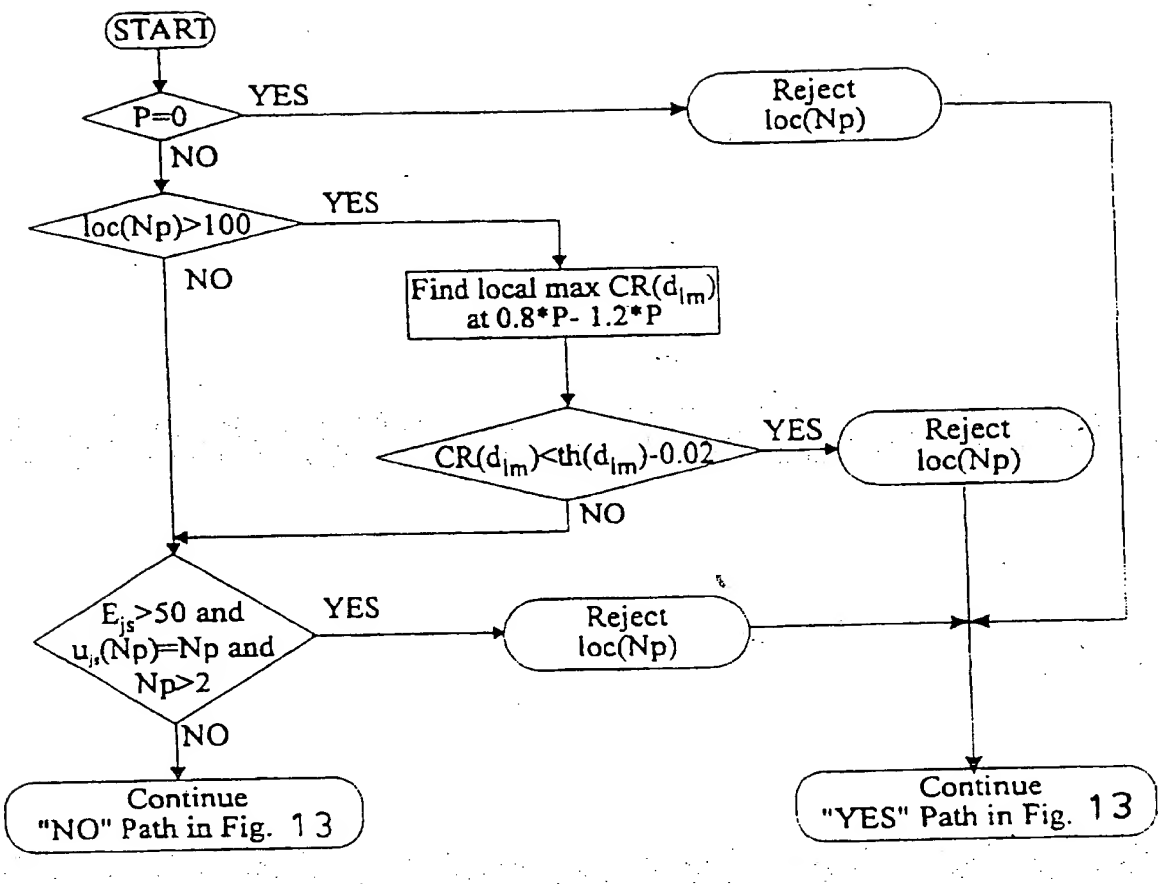

START
P=0
YES
NO
Reject loc(Np)
loc(Np)>100
YES
NO
Find local max CR(d_{lm}) at 0.8*P- 1.2*P
CR(d_{lm})<th(d_{lm})-0.02
YES
NO
Reject loc(Np)
E_{js}>50 and u_{js}(Np)=Np and Np>2
YES
NO
Reject loc(Np)
Continue "NO" Path in Fig. 13
Continue "YES" Path in Fig. 13

FIG. 14

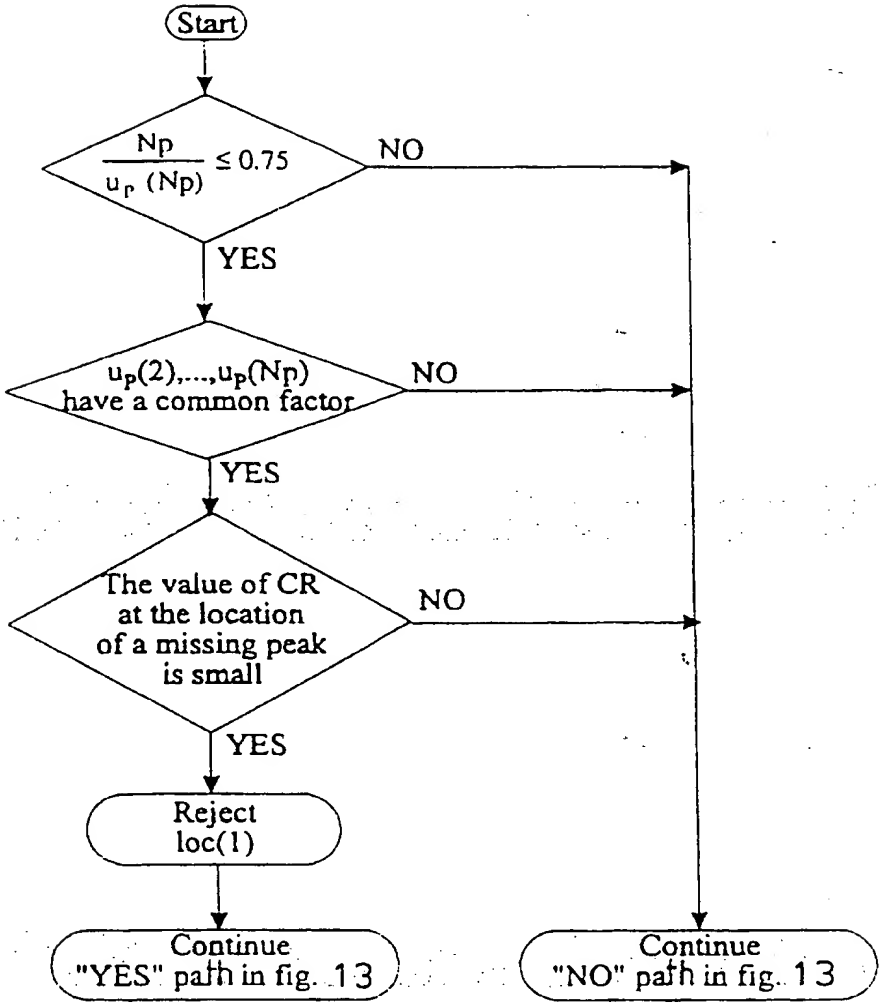

Start
Np / up (Np) ≤ 0.75
NO
YES
up(2),...,up(Np) have a common factor
NO
YES
The value of CR at the location of a missing peak is small
NO
YES
Reject loc(1)
Continue "YES" path in fig. 13
Continue "NO" path in fig. 13

FIG. 15

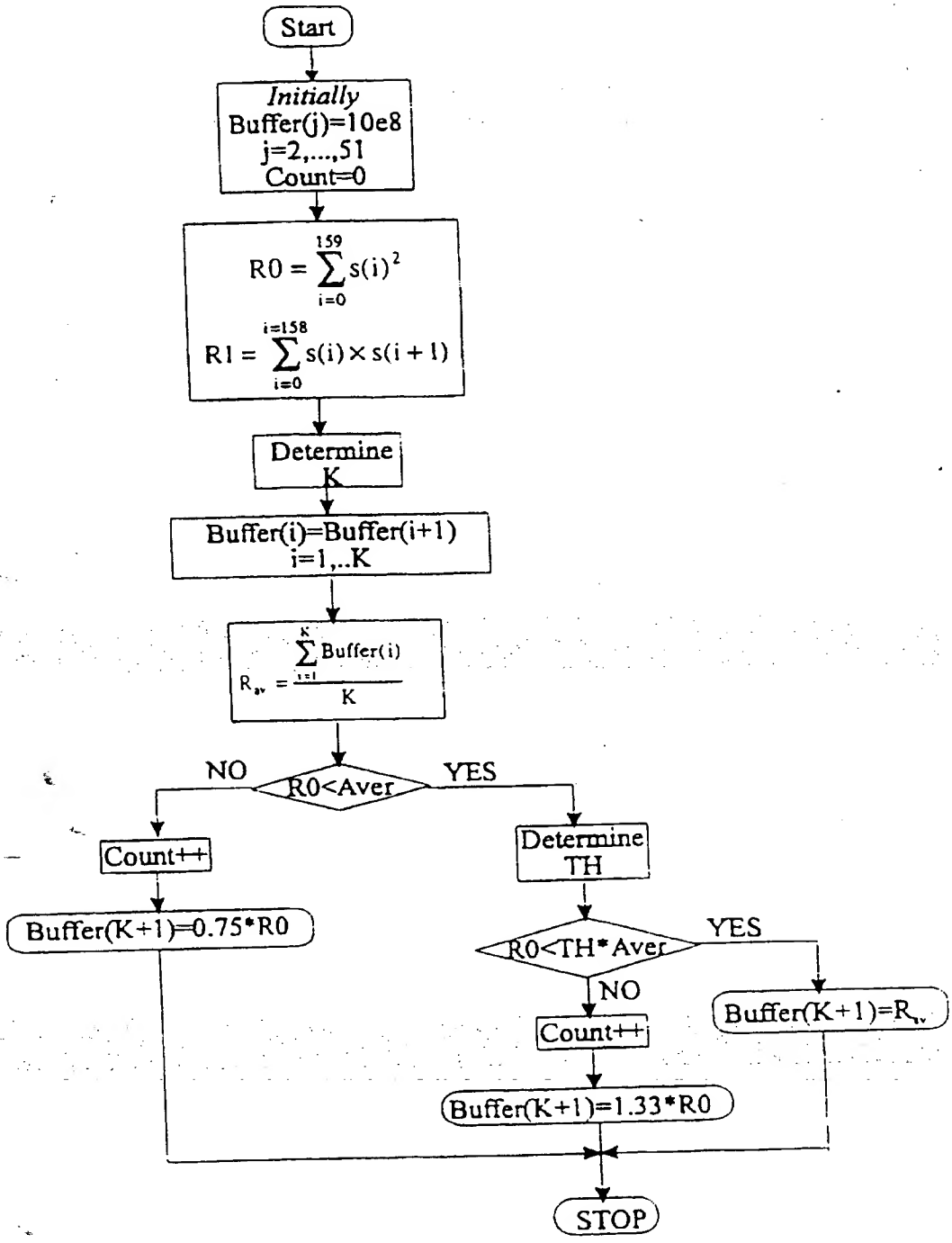

Start
Initially
Buffer(j)=10e8
j=2,...,51
Count=0
$R0 = \sum_{i=0}^{159} s(i)^2$
$R1 = \sum_{i=0}^{i=158} s(i) \times s(i+1)$
Determine K
Buffer(i)=Buffer(i+1)
i=1,..K
$R_{av} = \frac{\sum_{i=1}^{K} Buffer(i)}{K}$
NO
R0<Aver
YES
Count++
Buffer(K+1)=0.75*R0
Determine TH
R0<TH*Aver
YES
NO
Count++
Buffer(K+1)=1.33*R0
Buffer(K+1)=R_{av}
STOP

FIG. 16

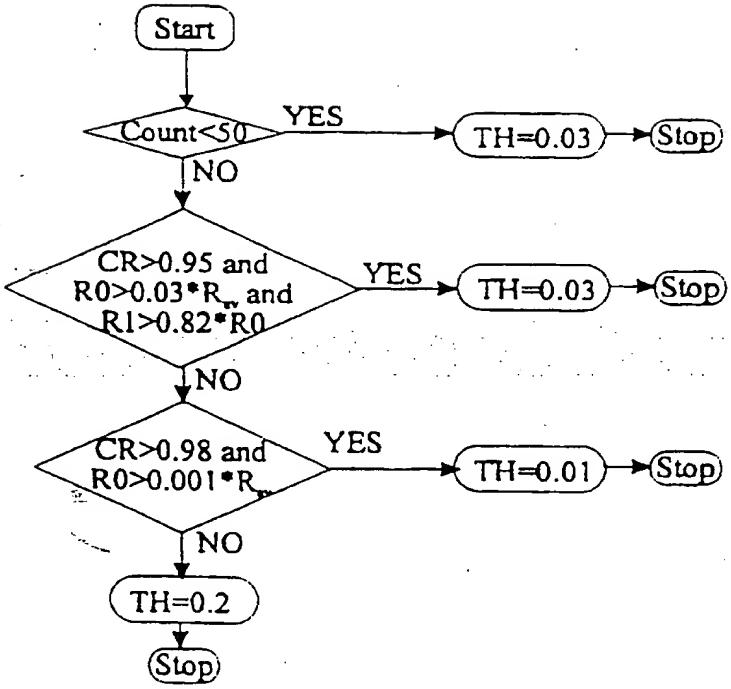

Start
Count<50
YES
TH=0.03
Stop
NO
CR>0.95 and
R0>0.03*R_w and
R1>0.82*R0
YES
TH=0.03
Stop
NO
CR>0.98 and
R0>0.001*R_w
YES
TH=0.01
Stop
NO
TH=0.2
Stop

FIG. 17

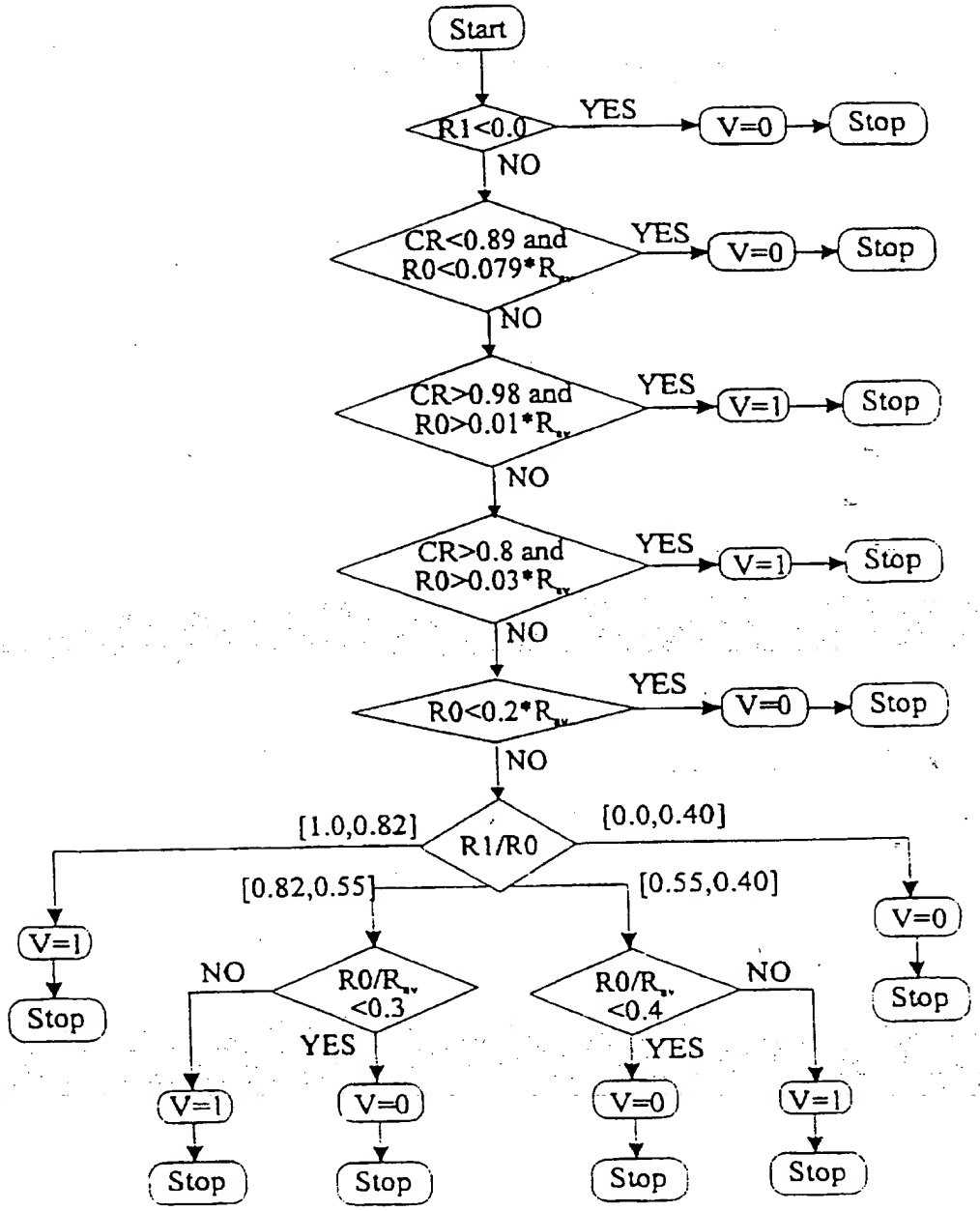
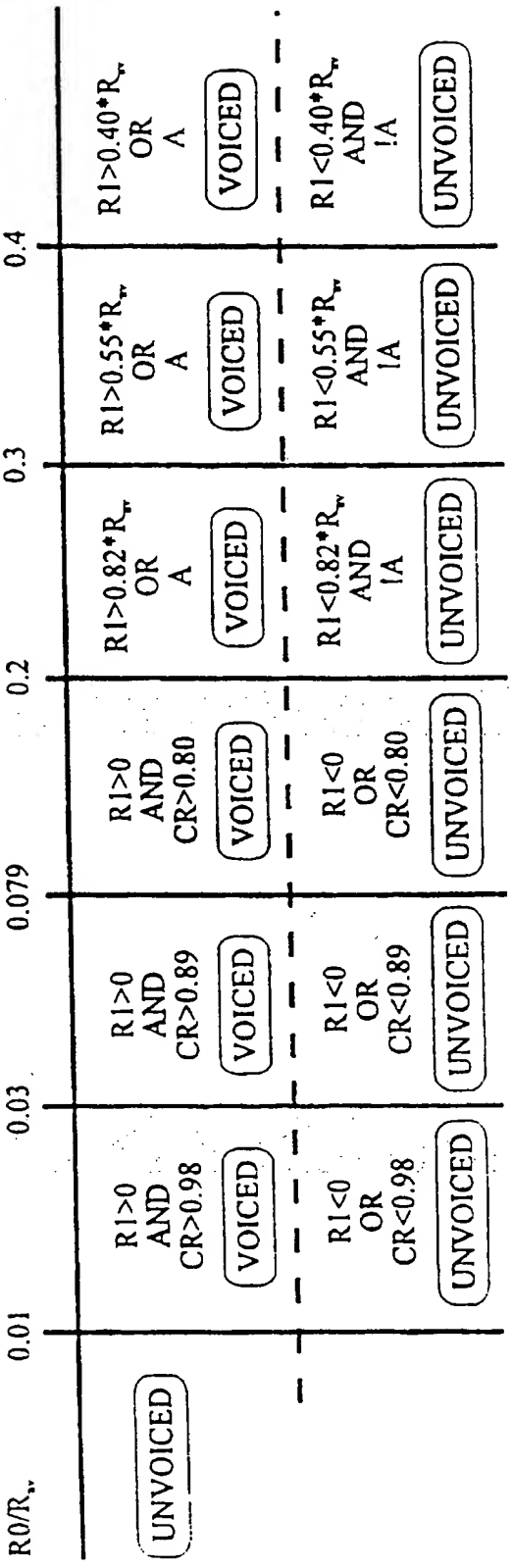

Start
R1<0.0
YES
V=0
Stop
NO
CR<0.89 and R0<0.079*R_w
YES
V=0
Stop
NO
CR>0.98 and R0>0.01*R_w
YES
V=1
Stop
NO
CR>0.8 and R0>0.03*R_w
YES
V=1
Stop
NO
R0<0.2*R_w
YES
V=0
Stop
NO
R1/R0
[1.0,0.82]
[0.82,0.55]
[0.55,0.40]
[0.0,0.40]
V=1
Stop
R0/R_w <0.3
NO
YES
V=1
Stop
V=0
Stop
R0/R_w <0.4
NO
YES
V=0
Stop
V=1
Stop
V=0
Stop

FIG.18



Condition A: R1>0 AND CR>0.80

Condition !A: R1<0 OR CR<0.80

FIG.19

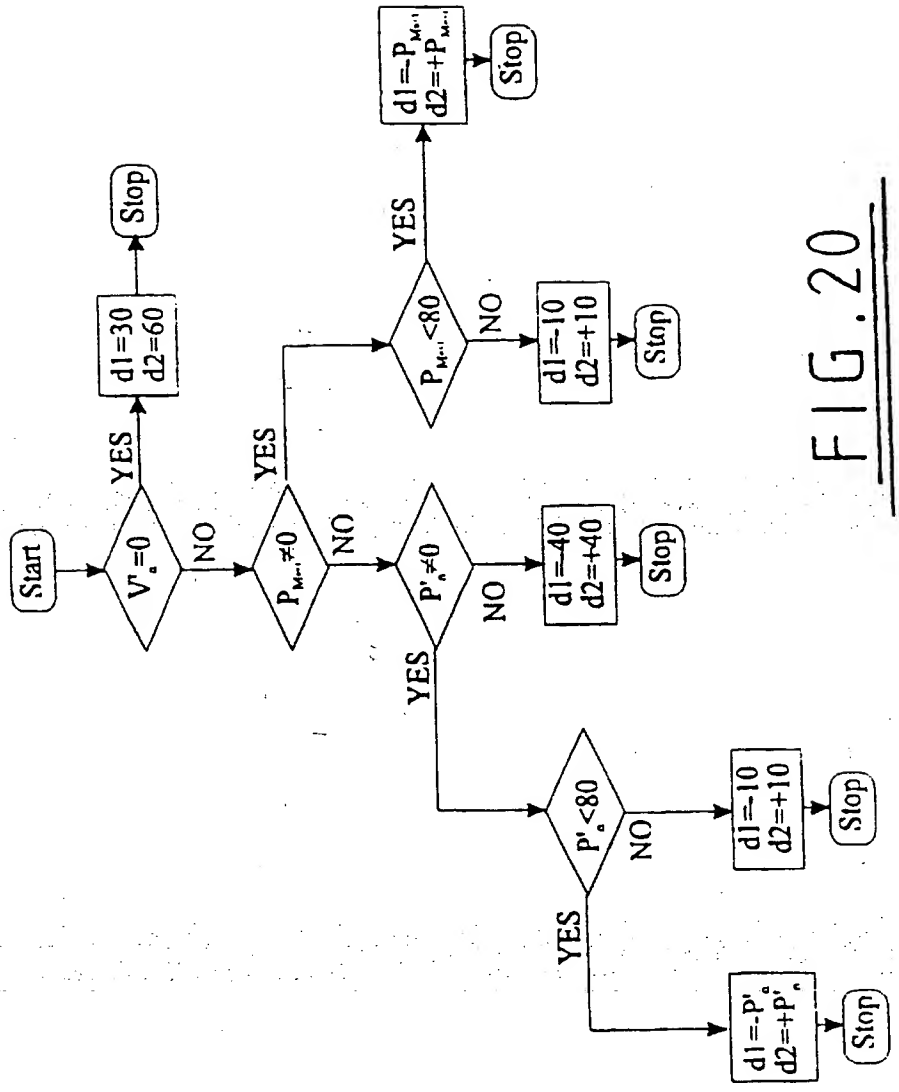

Start
V'a=0
YES
d1=30
d2=60
Stop
NO
PM...≠0
YES
PM...<80
YES
d1=-PM...
d2=+PM...
Stop
NO
d1=-10
d2=+10
Stop
NO
P'a≠0
NO
d1=-40
d2=+40
Stop
YES
P'a<80
NO
d1=-10
d2=+10
Stop
YES
d1=-P'a
d2=+P'a
Stop

FIG.20

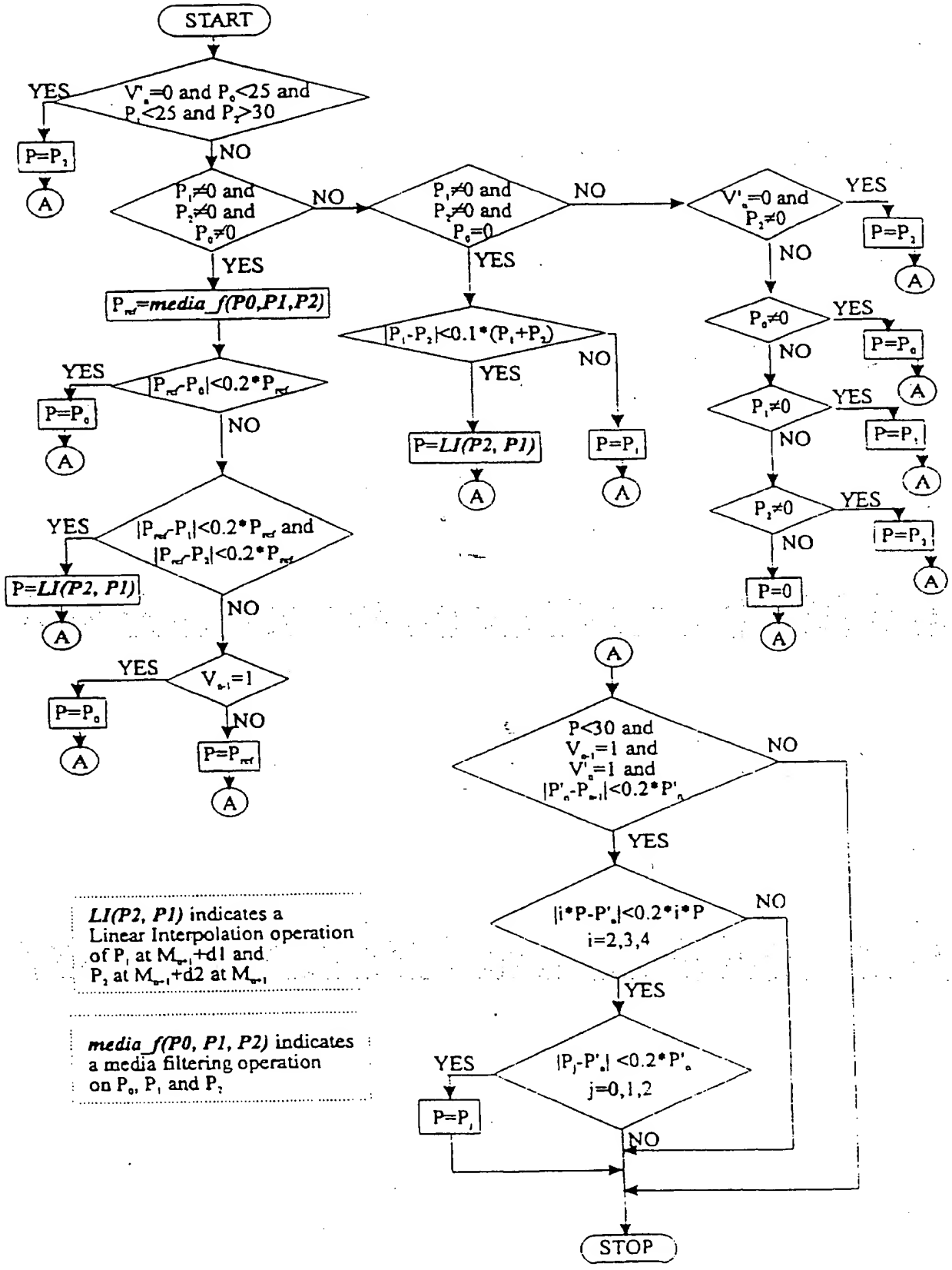

START
V'o=0 and Po<25 and P1<25 and P2>30
YES
NO
P=P2
A
P1≠0 and P2≠0 and P0≠0
NO
P1≠0 and P2≠0 and P0=0
NO
V'o=0 and P2≠0
YES
P=P2
A
NO
YES
Pref=media_f(P0,P1,P2)
|P1-P2|<0.1*(P1+P2)
YES
NO
P=LI(P2, P1)
P=P1
A
A
P0≠0
YES
P=P0
A
NO
P1≠0
YES
P=P1
A
NO
P2≠0
YES
P=P2
A
NO
P=0
A
|Pref-P0|<0.2*Pref
YES
P=P0
A
NO
|Pref-P1|<0.2*Pref and |Pref-P2|<0.2*Pref
YES
P=LI(P2, P1)
A
NO
Vn-1=1
YES
P=P0
A
NO
P=Pref
A
A
P<30 and Vn-1=1 and V'o=1 and |P'o-Pn-1|<0.2*P'o
NO
YES
|i*P-P'o|<0.2*i*P i=2,3,4
NO
YES
|Pj-P'o| <0.2*P'o j=0,1,2
YES
P=P1
NO
STOP
LI(P2, P1) indicates a Linear Interpolation operation of P1 at Mn-1+d1 and P2 at Mn-1+d2 at Mn-1
media_f(P0, P1, P2) indicates a media filtering operation on P0, P1 and P2

FIG. 21

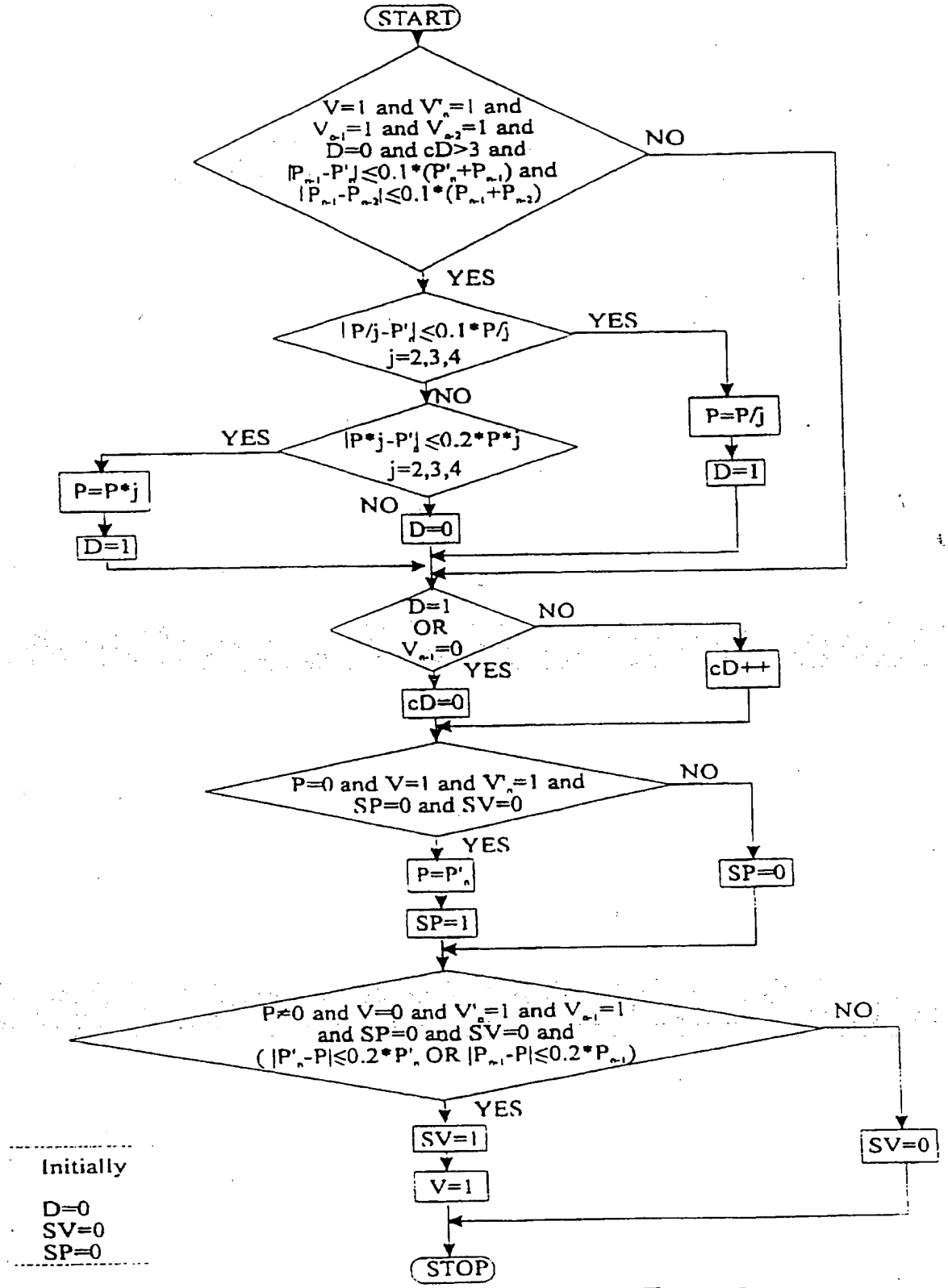
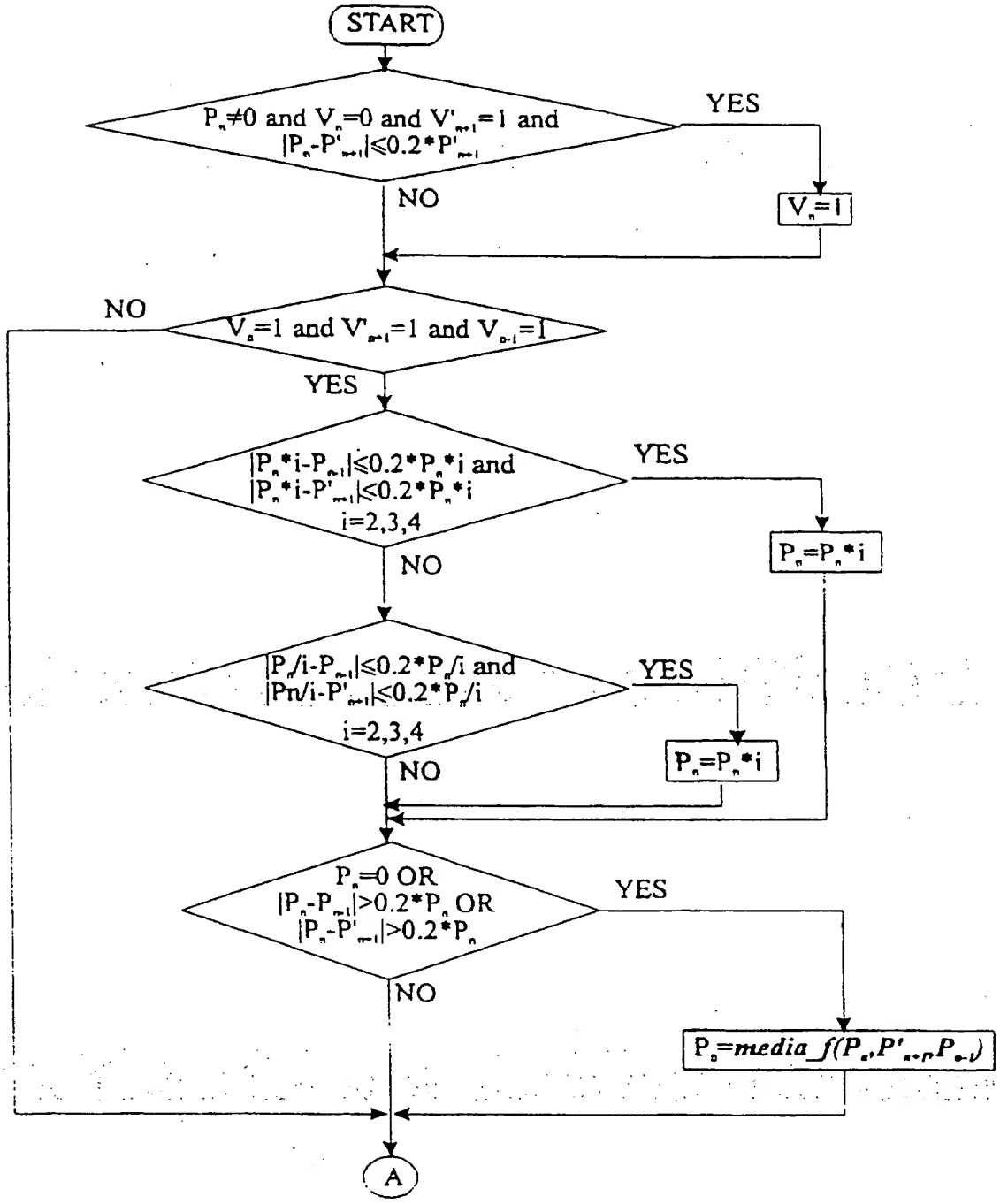

START
V=1 and V'n=1 and Vn-1=1 and Vn-2=1 and D=0 and cD>3 and |Pn-1-P'n|≤0.1*(P'n+Pn-1) and |Pn-1-Pn-2|≤0.1*(Pn-1+Pn-2)
NO
YES
|P/j-P'n|≤0.1*P/j j=2,3,4
YES
P=P/j
D=1
NO
|P*j-P'n|≤0.2*P*j j=2,3,4
YES
P=P*j
D=1
NO
D=0
D=1 OR Vn-1=0
NO
cD++
YES
cD=0
P=0 and V=1 and V'n=1 and SP=0 and SV=0
NO
SP=0
YES
P=P'n
SP=1
P≠0 and V=0 and V'n=1 and Vn-1=1 and SP=0 and SV=0 and (|P'n-P|≤0.2*P'n OR |Pn-1-P|≤0.2*Pn-1)
NO
SV=0
YES
SV=1
V=1
STOP
Initially
D=0
SV=0
SP=0

FIG. 22



media_f(P_n, P'_{n-1}, P_{n-1}) indicates a media filtering operation on P_n, P'_{n-1} and P_{n-1}

FIG. 23

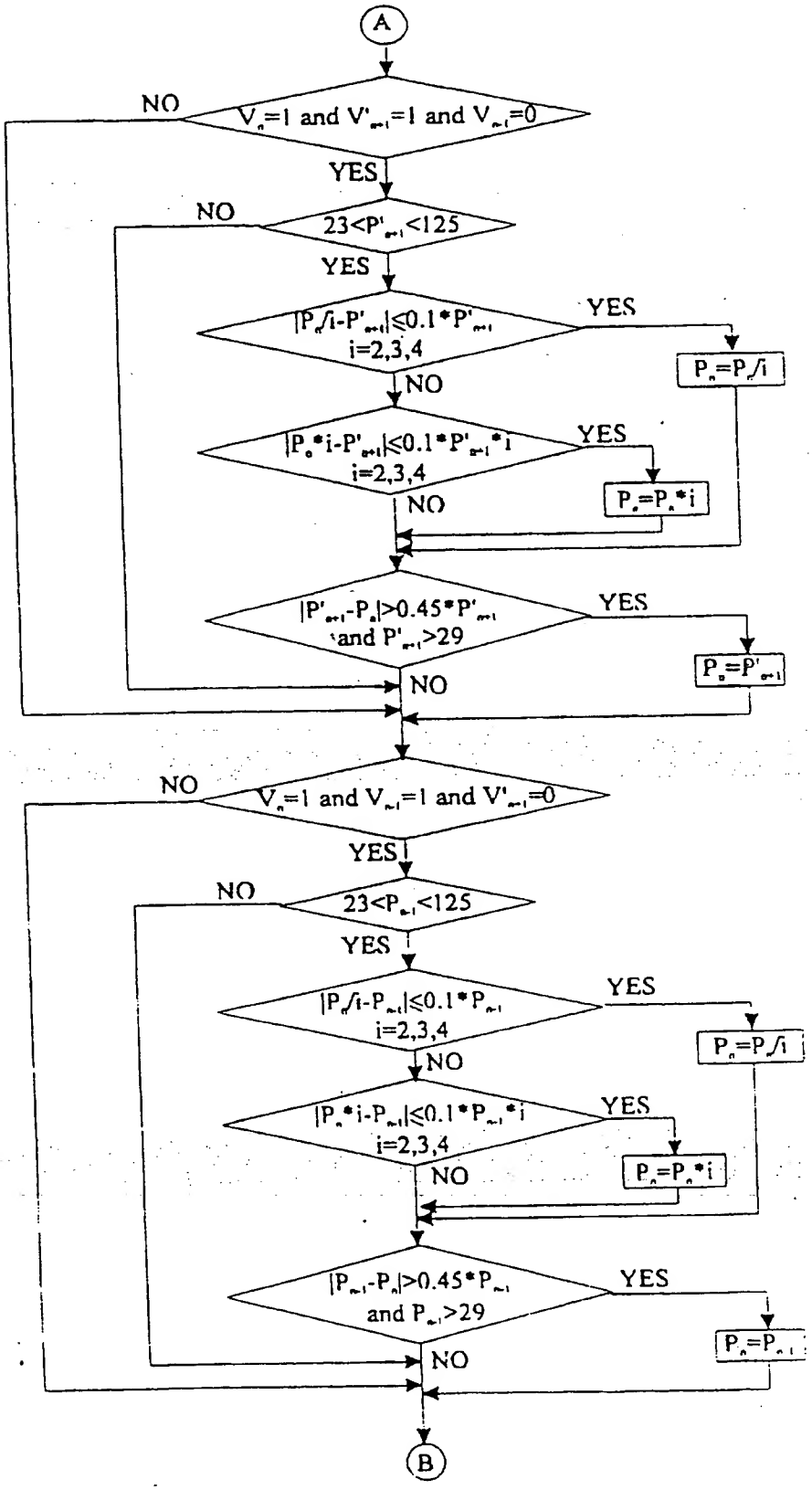

A
V_n=1 and V'_{n-1}=1 and V_{n-1}=0
NO
YES
23<P'_{n-1}<125
NO
YES
$|P_n/i-P'_{n-1}|\leq 0.1*P'_{n-1}$ i=2,3,4
YES
$P_n=P_n/i$
NO
$|P_n*i-P'_{n-1}|\leq 0.1*P'_{n-1}*i$ i=2,3,4
YES
$P_n=P_n*i$
NO
$|P'_{n-1}-P_n|>0.45*P'_{n-1}$ and $P'_{n-1}>29$
YES
$P_n=P'_{n-1}$
NO
V_n=1 and V_{n-1}=1 and V'_{n-1}=0
NO
YES
23<P_{n-1}<125
NO
YES
$|P_n/i-P_{n-1}|\leq 0.1*P_{n-1}$ i=2,3,4
YES
$P_n=P_n/i$
NO
$|P_n*i-P_{n-1}|\leq 0.1*P_{n-1}*i$ i=2,3,4
YES
$P_n=P_n*i$
NO
$|P_{n-1}-P_n|>0.45*P_{n-1}$ and $P_{n-1}>29$
YES
$P_n=P_{n-1}$
NO
B

FIG.24

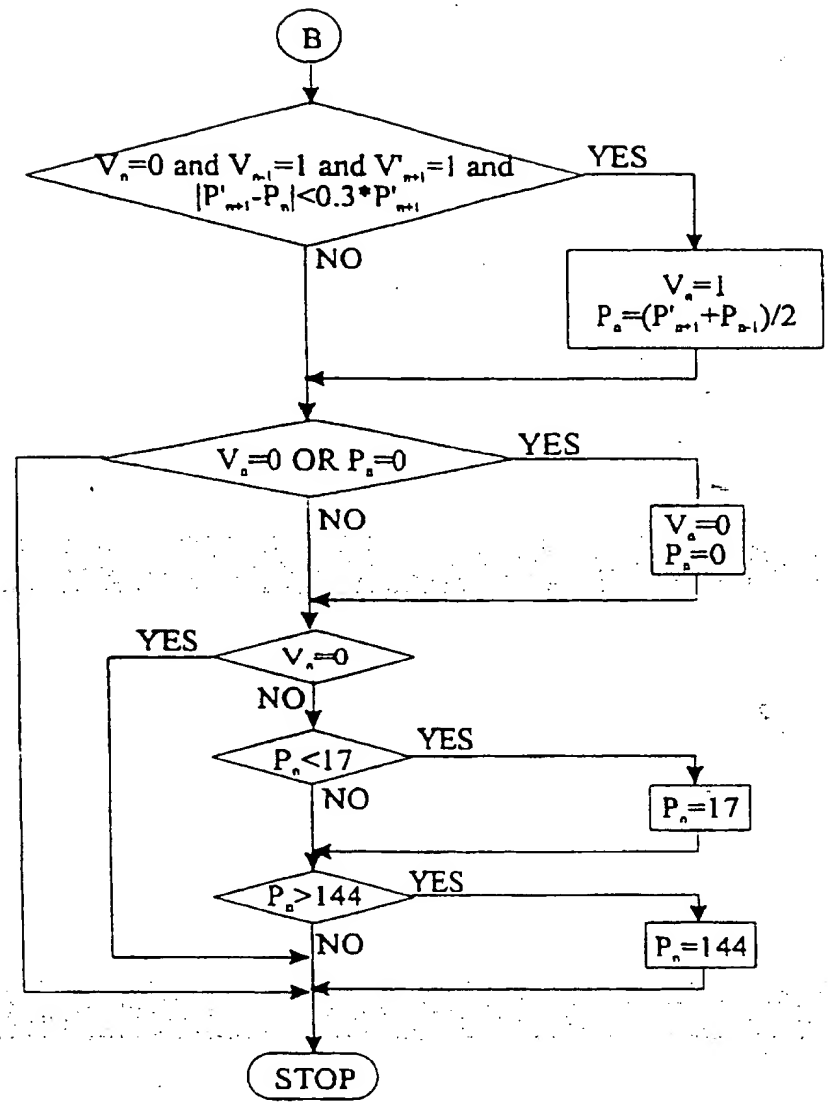

B
V_n=0 and V_{n-1}=1 and V'_{n+1}=1 and $|P'_{n+1}-P_n|<0.3*P'_{n+1}$
YES
NO
V_n=1
$P_n=(P'_{n+1}+P_{n-1})/2$
V_n=0 OR P_n=0
YES
NO
V_n=0
P_n=0
V_n=0
YES
NO
P_n<17
YES
NO
P_n=17
P_n>144
YES
NO
P_n=144
STOP

FIG.25

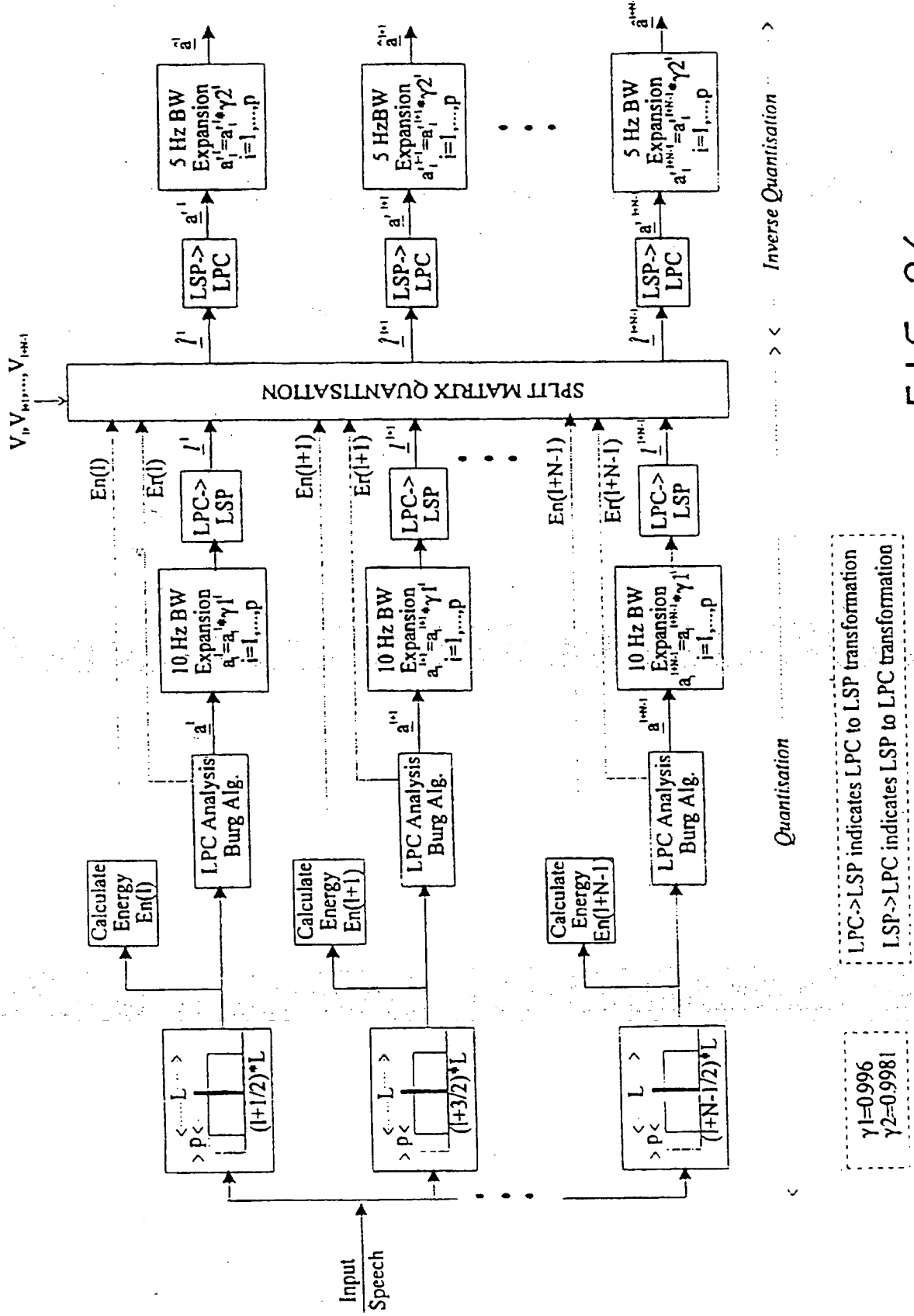

Input Speech
Calculate Energy En(l)
LPC Analysis Burg Alg.
10 Hz BW Expansion
LPC-> LSP
SPLIT MATRIX QUANTISATION
LSP-> LPC
5 Hz BW Expansion
Calculate Energy En(l+1)
Calculate Energy En(l+N-1)
Quantisation
Inverse Quantisation
LPC->LSP indicates LPC to LSP transformation
LSP->LPC indicates LSP to LPC transformation
γ1=0.996
γ2=0.9981

FIG. 26

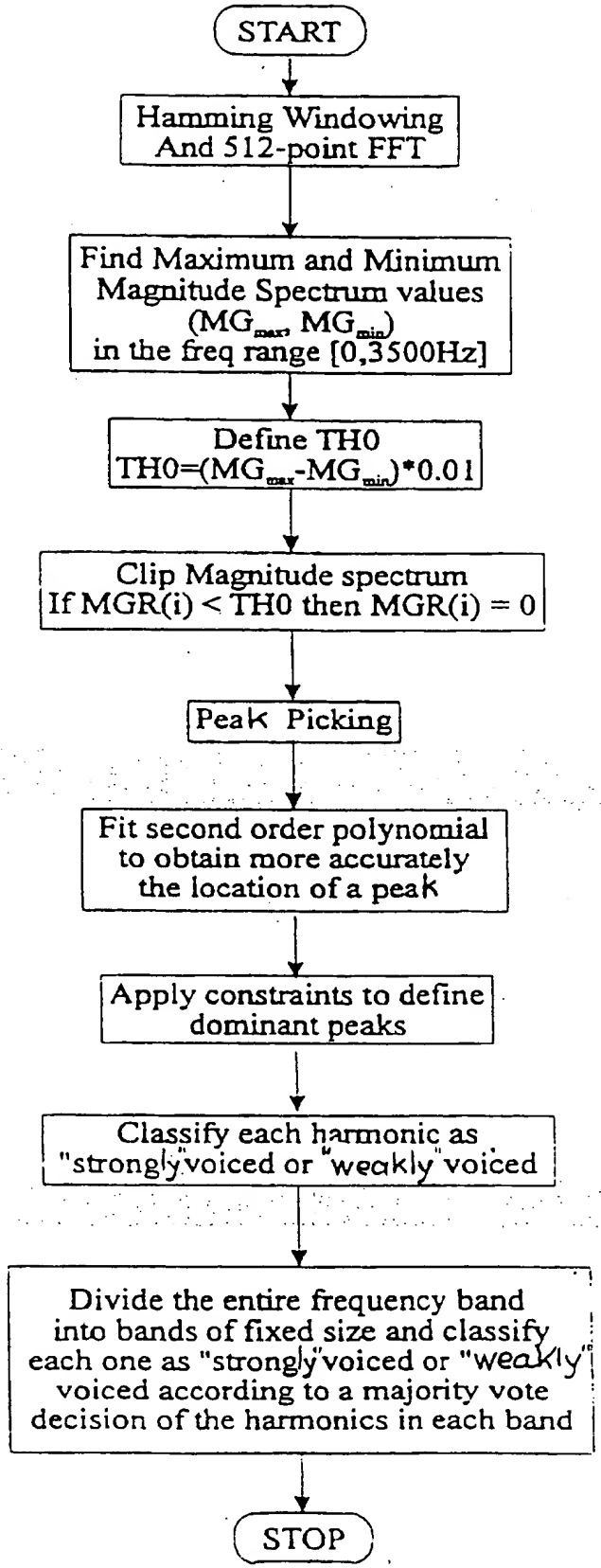

START
Hamming Windowing And 512-point FFT
Find Maximum and Minimum Magnitude Spectrum values (MG_{max}, MG_{min}) in the freq range [0,3500Hz]
Define TH0 TH0=(MG_{max}-MG_{min})*0.01
Clip Magnitude spectrum If MGR(i) < TH0 then MGR(i) = 0
Peak Picking
Fit second order polynomial to obtain more accurately the location of a peak
Apply constraints to define dominant peaks
Classify each harmonic as "strongly"voiced or "weakly"voiced
Divide the entire frequency band into bands of fixed size and classify each one as "strongly"voiced or "weakly" voiced according to a majority vote decision of the harmonics in each band
STOP

FIG. 27

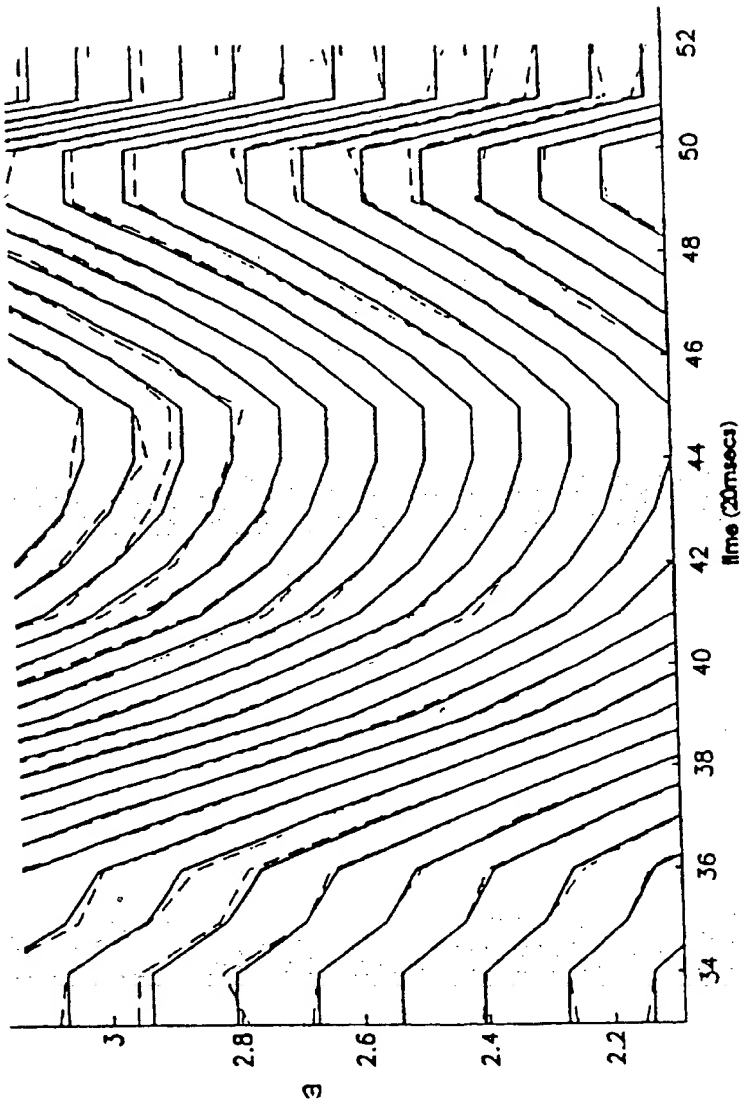


FIG. 31

Figure 3.13

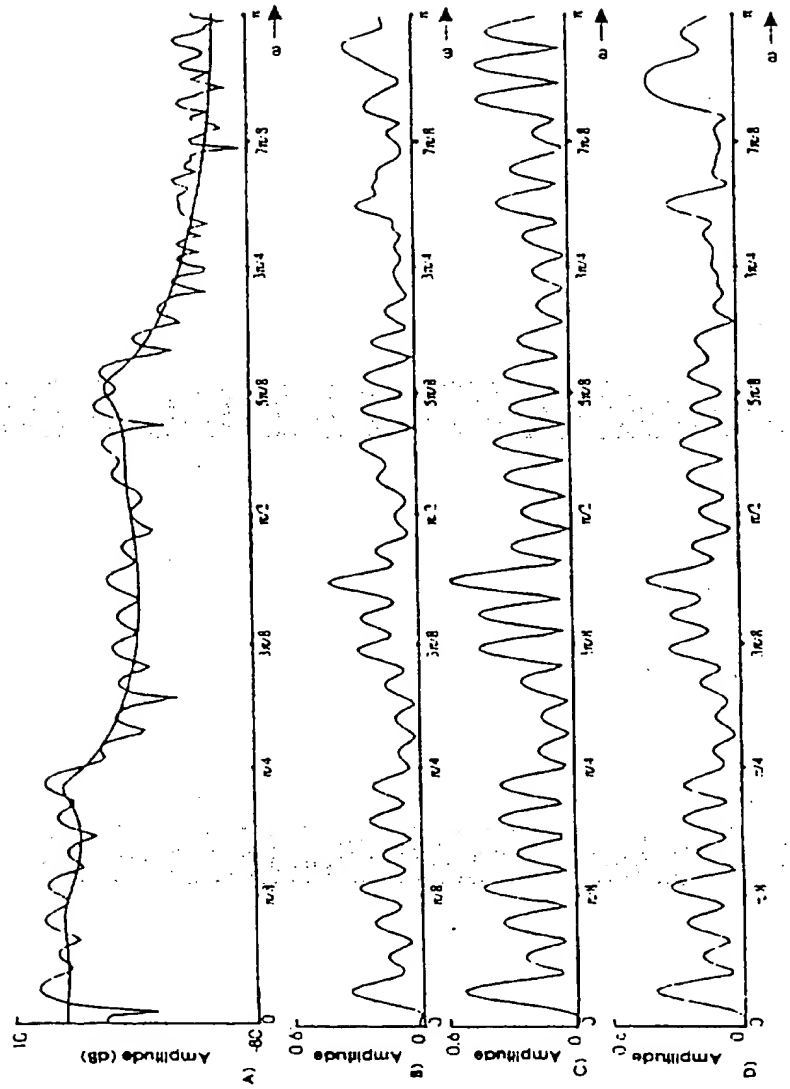


FIG. 32

Figure 3.15

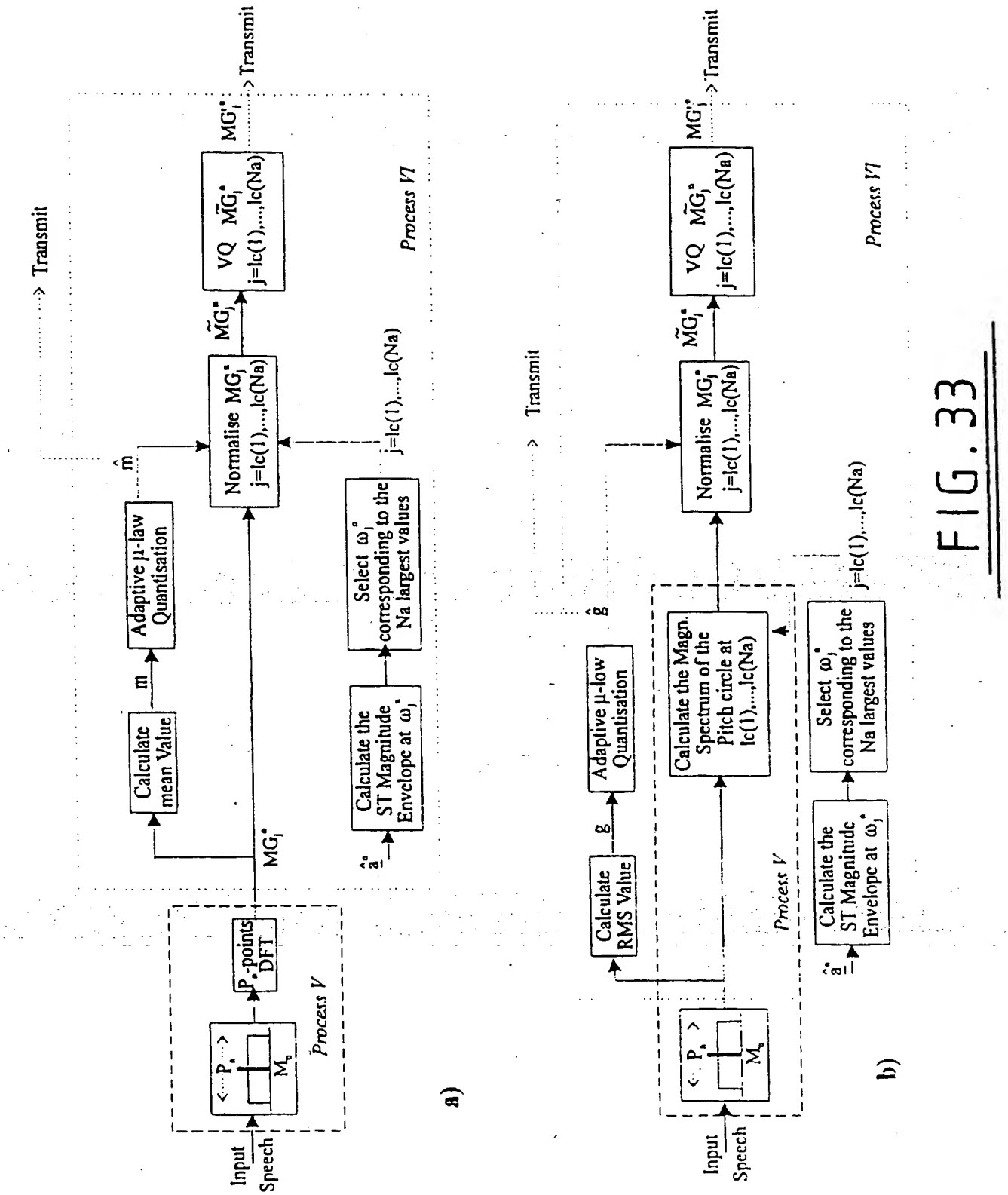

Input Speech
Process V
Calculate mean Value
Adaptive μ-law Quantisation
Normalise MG_j^* j=Ic(1),...,Ic(Na)
VQ MG_j^* j=Ic(1),...,Ic(Na)
Transmit
Process VI
Calculate the ST Magnitude Envelope at ω_j^*
Select ω_j^* corresponding to the Na largest values
a)
Calculate RMS Value
Adaptive μ-low Quantisation
Calculate the Magn. Spectrum of the Pitch circle at Ic(1),...,Ic(Na)
b)

FIG. 33

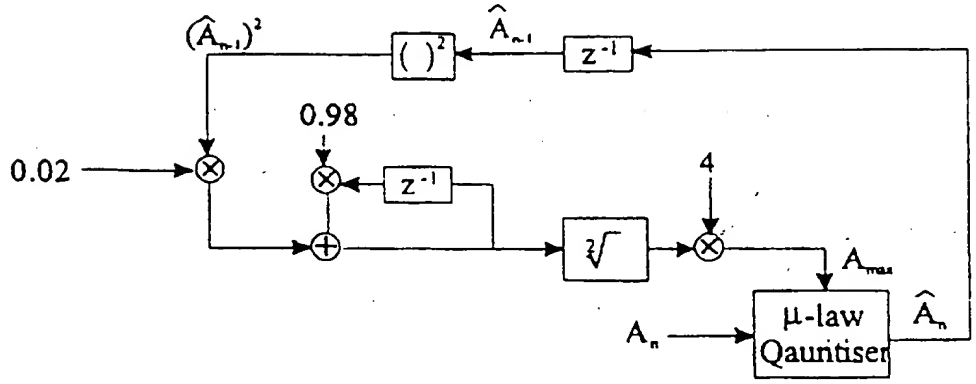


FIG.34

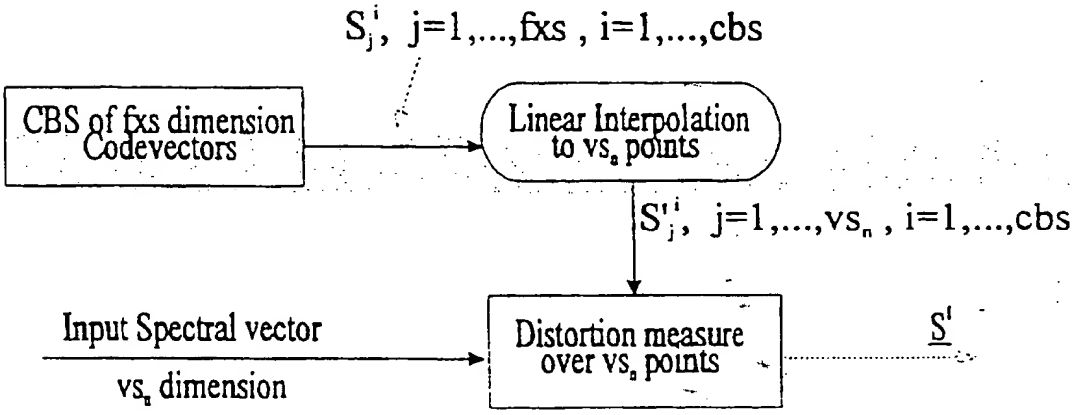


FIG.35

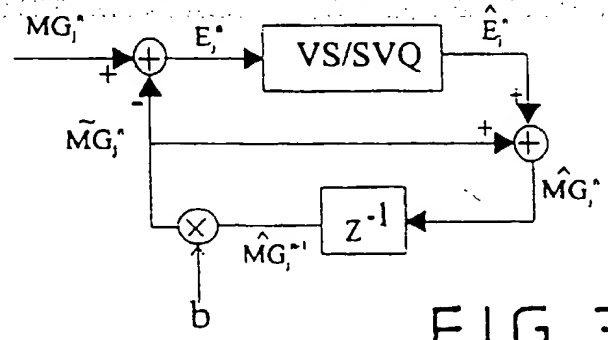


FIG.36

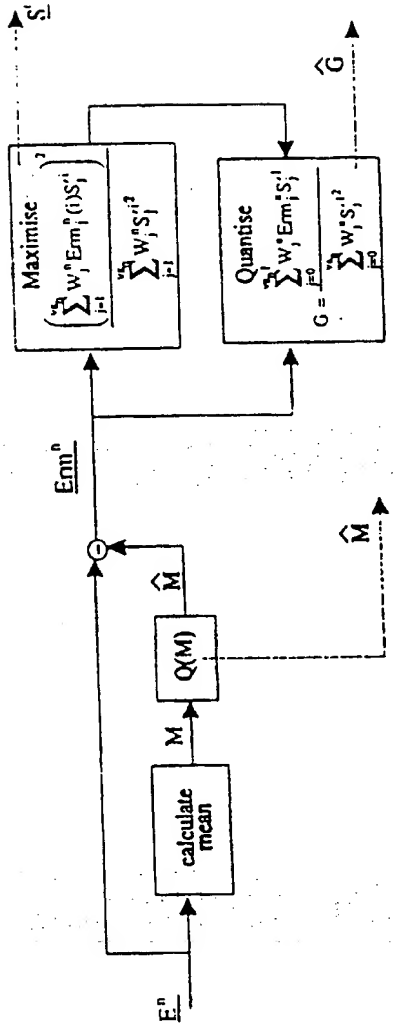


FIG.37

INTERNATIONAL SEARCH REPORT

Internat Application No PCT/GB 97/01831

A. CLASSIFICATION OF SUBJECT MATTER

G 10 L 5/02,G 10 L 3/00,G 10 L 3/02,G 10 L 7/02

According to International Patent Classification (IPC) or to both national classification and IPC 6

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G 10 L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X,P	WO 96/27870 A1 (BRITISH TELECOMMUNICATIONS PUBLIC LIMITED COMPANY) 12 September 1996 (12.09.96), fig. 1-3, abstract, claims 1-4, page 1, line 25 - page 4, line 4. --	1
A	EP 0703565 A2 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 27 March 1996 (27.03.96), fig. 1-6, abstract, claims 1-12, page 4, line 7 - page 5, line 58. --	1-47
A	EP 0490740 A1 (THOMSON-CSF) 17 June 1992 (17.06.92),	1-47

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search
06 October 1997

Date of mailing of the international search report
12. 11. 97

Name and mailing address of the ISA
European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+ 31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+ 31-70) 340-3016

Authorized officer
BERGER e.h.

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/GB 97/01831

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	fig. 1-5, abstract, claims 1-6. ----	

ANHANG

zum internationalen Recherchenbericht über die internationale Patentanmeldung Nr.

ANNEX

to the International Search Report to the International Patent Application No.

ANNEXE

au rapport de recherche international relatif à la demande de brevet international n°

PCT/GB 97/01831 SAE 165942

In diesem Anhang sind die Mitglieder der Patentfamilien der im obengenannten internationalen Recherchenbericht angeführten Patentdokumente angegeben. Diese Angaben dienen nur zur Unterrichtung und erfolgen ohne Gewähr.

This Annex lists the patent family members relating to the patent documents cited in the above-mentioned international search report. The Office is in no way liable for these particulars which are given merely for the purpose of information.

La présente annexe indique les membres de la famille de brevets relatifs aux documents de brevets cités dans le rapport de recherche international visée ci-dessus. Les renseignements fournis sont donnés à titre indicatif et n'engagent pas la responsibilité de l'Office.

In Recherchenbericht angeführtes Patentdokument Patent document cited in search report Document de brevet cité dans le rapport de recherche	Datum der Veröffentlichung Publication date Date de publication	Mitglied(er) der Patentfamilie Patent family member(s) Membre(s) de la famille de brevets	Datum der Veröffentlichung Publication date Date de publication
WO A1 9627870	12-09-96	AU A1 49488/96	23-09-96
EP A2 703565	27-03-96	JP A2 8095589 US A 5671330	12-04-96 23-09-97
EP A1 490740	17-06-92	CA AA 2057139 FR A1 2670313 US A 5313551	12-06-92 12-06-92 17-05-94